Analytical strategies in Imaging Genetics: Assessment of potential risk factors for neurodevelopmental domains

Natàlia Vilor Tejedor

TESI DOCTORAL UPF / 2018

DIRECTORS DE LA TESI

Dr. Juan Ramón González Ruiz

Co-Director: Prof. Jordi Sunyer Deu

Barcelona Institute for Global Health (ISGlobal)
Department of Experimental and Health Sciences

This Predoctoral Research has been performed thanks to the following subsidized projects:

- Agència de Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya – Fons Social Europeu (2015 FI_B 00636; 2016 FI_B 00636; 2017 FI_B 00636).

- European Research Council under the ERC Grant Agreement (ERC-AdG 2010 GA268479)

- Ministerio de Economía e Innovación (MTM2015-68140-R, MTM2012-38067-C02-02)

and other entitites:

Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Open Multiscale Systems Medicine Cost Action 1520 (OpenMultiMed). Red Nacional de Bioestadística (BIOSTATNET), and Sociedad Española de Biometría.

"The journey is always more important than the destination"

- Jeff Klein

Acknowledgements

L'ànima està feta d'un material tan misteriosament elàstic que un sol esdeveniment pot engrandir-la prou per contenir l'infinit¨

Volia començar els agraïments per la Dr. Malu Calle, qui em va animar a iniciar aquest camí. Malu, tu has sigut qui vas provocar l'*esdeveniment* que m'ha portat fins aquí. Gràcies, per obrir-me les portes d'aquest món, donar-me els ànims, compartir els teus coneixements i recolzar-me per continuar. Gràcies per tot el temps, per totes les oportunitats, pel teu suport i per la teva confiança incondicional. Ets una de les persones que més admiro, i com la meva *mare* dins del món de la recerca. Probablement, molt bona part de la persona que sóc ara, és gràcies a tu. Començo una nova etapa en la que espero, ara si, poder formar-me i gaudir al teu costat. Gràcies de tot cor per tot.

Silvi!!!!!! i probablement no hagués acabat mai si no hagués estat pel teu suport i *coaching*! M'has vist créixer, evolucionar, m'has vist caure i tornar-me a aixecar. M'has guiat, m'has escoltat, has sigut *LA* meva supervisora, una confessora, una font de carinyo, una menja hamburgueses (fins que vam descobrir el "xino" més bo de Barcelona), una bona amiga i com una germana. Ha estat un plaer poder "aguantar" a la Dr. Alemany tot aquest temps, no dubtis que ho tornaria a fer (i la meva bessona segur que també ;P). I no continuu perque ja saps la frase no políticament correcta que vindria després... Mil gràcies per tot, una part important d'aquesta tesi és molt teva també.

Gràcies també al meu director de tesis, el Prof. Jordi Sunyer, per l'entusiasme, i per saber deixar-me fer allò que m'agrada fer. Jordi, ets un exemple de saviesa i d'amor per la recerca.

Agraeixo molt totes les teves paraules assenyades i accions amables. Gràcies per guiar-me i donar-me l'oportunitat de madurar com investigadora.
Gracias a mi otro co-director, Juan Ramón González, por darme la oportunidad de poder trabajar en esta institución y en esta línia de investigación. Gracias por tu entusiasmo.
Gracias Alejandro, por tan buenas charlas, consejos, comentarios y correciones. He tenido mucha suerte de tener una persona como tú tan cerca. Mucho de lo que aquí se presenta también te lo debo a ti.
Gràcies també a la Mariona Bustamante per la seva energia, i honestedat. Mariona, gràcies pel teu temps, dedicació i acollida. Si en una persona penso al parlar de professionalitat, ets tu.
Gràcies també a totes les gestores que amb el vostre granet de sorra heu contribuït a agilitzar-me i facilitar-me el meu temps durant el doctorat. En especial les que heu treballat directament amb mi, les meves estimades: Iolanda, Gemma Punyet, Maria Elena, Anna Sillero i Mònica Millán. Però també a totes les altres. No sabeu quant d'importants sou en tot aquests resultats i en tota la nostra investigació, i sobretot quant d'importants heu estat per mi!
Gràcies també a tots els tècnics, project managers, data managers i personal que formeu part d'aquesta institució o que n'heu format. Sou grans professionals, però millors persones. Sense tota la vostra feina no faríem res. En especial, gràcies tete Albert, pels consells, per animar-me, pels riures i l'alegria infinita. Ets la felicitat feta persona i de les persones amb més bon cor que conec.
Gracias tambien a mi super griega favorita, Dania, por ser tan auténtica, por todos los consejos; por la luz de Syros al atardecer y la oscuridad de Boston al anochecer. Eres un sol de persona y un gran ejemplo a seguir.

Gràcies al grup de predocs, sobretot aquells que per coincidència espai-temps hem compartit diferents experiències/viatges/cerveses i molt més. Gràcies Ione, per l'alegria, i les confidències. Gràcies Ita, em queden molts bons records i molts riures de l'aventura londinenc (ningú sap encara com vam aconseguir agafar el vol...). Gracias también Cyntia y Natalie por esa "linia" predoc en la sala B, i també Carles i Marcos per compartir de principi a fi aquesta aventura/etapa bioinformàtica. Gràcies també companys de la sala B (que encara resistiu o que vàreu marxar). Gràcies Marta, Esther, Ignasi, Lucas, David M., David D., Mikel, Elisa, Gascón.

Gracias al grupo de BRGE-eros por vuestro frikismo exponencial. També al grup de "neuro" per acollir-me com una més, i a tota la gent del Childhealth program.

Gràcies al grup de Neuroimatge de l´Hospital del Mar, en especial al Dr. Jesús Pujol per tants bons feedbacks i tardes debatent sobre estructura i funcionalitat del cervell.

Gracias también a todos los miembros de la Sociedad Española de Biometría y a la red Nacional de Bioestadística (BIOSTATNET). En especial a mi grupo de jóvenes preferido: Marta, Moises, Urko, Núria, Altea, Pilar, Manuel, Maider, Josu, Irantzu, Blanca. Y también al de más veteranos; Lupe, David C., Vicente, Carmen, Pere, Anna Espinal, Inma, Arantxa. Guardo fantásticos recuerdos de esta etapa con todos vosotros. Gracias por vuestro trabajo, vuestra constancia, motivación y ayuda. Este camino no hubiera sido el mismo sin vosotros.

Mila esker, baita ere, Matematika Aplikatua, Estatistika eta Ikerkuntza Operatiboa Sailari. Bereziki, nire esker onenak eman nahi dizkiet Inmaculada Arosteguiri, Irantzu Barriori eta Arantza Urkaregiri. Hasieratik izandako konfiantzagatik, nirekin izandako adeitasunagatik, Bilbon egindako

harreragatik eta zuen denbora eta aholku guztiengatik. Pertsona bikainak zarete eta balio bikainak helarazten dituzue. Baita "kamiseta urdinak" ere, nire PhD-bihotzeko txoko txiki batean izango zaituztet beti.
Many thanks to all the people from the Rotterdam Study and Erasmus Medical Center, specially to Hieab and Genna. Also thanks to Prof. M. Arfan Ikram for your support and for giving me the opportunity to learn about Imaging Genetics and take part of your research team.

Gràcies també a tots aquells, que externs i durant aquesta aventura m'heu fet costat en els moments durs; A mi flamante bigTwin, porque eres como un hermano y confidente. A les meves matemàtiques solteres, casades o nouvingudes; Noe, Jana, Anita, Patatilla Jr., Nadia, Abril, Pati, Laura, Carme, i als respectius consortes. Al meu catalano-francés reconvertit a casteller favorit!!! Al trio de pivonatchys más divino del mundo mundial; Flax, Eli, Gemma que bo tenir-vos! Gràcies a la resta de Pivonatchys, i a la petita Isuchi per ser tan original amb el nom ;)! Gracias Ana, por ayudarme a descrubrir una nueva versión de Natàlia. Gràcies a aquelles Plahaies i consortes que em vau adoptar de cor, i que continueu preocupant-vos i caminant al meu costat. També agrair a l'equip de futbol sala de la Pompeu per l´acollida, "*Rojo-Locura, que tenemos aguante de eso no hay duda*". Als kimiko-molonguis exploradors de món i nits de gominoles... *aaaarriba*! A mi sister por quedarte tú con el gen "*Vilor*". Als barberencs i barberenques de la EAB per ensenyar-me una mica de política i molt de saber fer. A la Belén per totes les nits de mojitos, just dance i sushi casero (i als porcs senglars per deixar-nos sobreviure). A mi Esteve, por tus aventuras locas, por tu cálido hogar en Islandia y por todas las anécdotas vividas.

Muito obrigada a todas as pessoas maravilhosas que conheci no Brasil, i arreu del món. A las noches de tango, samba, de caipirinhas i de festes majors. A tots els que segur em descuido però que heu aportat el vostre granet d´alegria durant aquesta etapa. Fins i tot als que vau marxar a temps per donar-me l'oportunitat de descobrir totes les meravelles que m'esperaven! A tots, tots, moltíssimes gràcies!

I gràcies Marina, per fer-me descobrir de nou la realitat de la felicitat; "*I've been through it all and nothing works better than to have someone you love hold you*".

"La vie ne mérite pas qu'on se préocupe autant"
– *Marie Curie*

SUMMARY

Imaging Genetics (IG) aims to test how genetic information influences brain structure and function, cognitive processes and complex neurodevelopmental domains, combining magnetic resonance imaging-based brain features and genetic data from the same individual. IG studies represent an opportunity to deepen our knowledge of the biological mechanisms of neurodevelopmental domains and complex brain disorders.

Most studies focus on individual correlation and association tests between a subset of genetic variants (usually single nucleotide polymorphisms, SNPs) and a single measurement of the brain. Despite the great success of univariate approaches, given the current focus of imaging genetic studies in which genome-wide, whole-brain studies should be analyzed, the development of novel statistical methods becomes crucial.

The main aim of this thesis consists of investigating genetic determinants of structural brain change, which in turn affect neurodevelopmental domains. We propose the application and development of statistical strategies to improve the assessment of significant relationships associated with neurodevelopmental domains. Specifically, we focus our research efforts on understanding what genomic changes in the cerebral structure allow improvements in the assessment of risk factors associated with Attention-Deficit/Hyperactivity disorder domains, and related cognitive processes such as attention function.

RESUM

Els estudis que combinen la informació genètica i de neuroimatge (IG) pretenen provar com la informació genètica influeix en l'estructura i funció cerebral, en el comportament, i en els dominis del neurodesenvolupament, combinant la informació extreta de ressonàncies magnètiques del cervell i de la informació genètica d'un mateix individu. Els estudis d'IG representen una oportunitat per aprofundir en el coneixement dels mecanismes biològics dels dominis del desenvolupament neurològic.

La majoria dels estudis es centren en la correlació individual i en proves d'associació entre un subconjunt de variants genètiques (en general polimorfismes d'un únic nucleòtid, SNPs) i una única mesura d'una regió cerebral. Però, malgrat el gran èxit en l'enfocament univariat, donades les perspectives actuals dels estudis d'IG, en els quals es pretenen analitzar les relacions cerebrals de tot el genoma envers tota la informació del cervell, el desenvolupament de nous mètodes estadístics específics esdevé crucial.

L'objectiu principal d'aquesta tesi consisteix a investigar els determinants genètics relacionats amb els canvis estructurals del cervell, que a la vegada, afecten els dominis del neurodesenvolupament. Proposem l'aplicació i el desenvolupament d'estratègies estadístiques per millorar l'avaluació de les relacions biològiques associades als dominis del neurodesenvolupament. Específicament, centrem els nostres esforços de recerca en comprendre quins canvis genètics que influeixen l'estructura cerebral permeten millorar l'avaluació dels factors de risc associats als dominis del trastorn per dèficit d'atenció i hiperactivitat, i a processos cognitius relacionats, com la funció d'atenció.

RESUMEN

Los estudios de genetica y neuroimagen (IG) pretende probar cómo la información genética influye en la estructura y función cerebral, el comportamiento y los dominios del neurodesarrollo, combinando la información extraida de resonancias magnéticas del cerebro y de datos genéticos del mismo individuo. Los estudios de IG representan una oportunidad para profundizar nuestro conocimiento de los mecanismos biológicos de los dominios del desarrollo neurológico.

La mayoría de los estudios se centran en la correlación individual y en pruebas de asociación entre un subconjunto de variantes genéticas (por lo general polimorfismos de un único nucleótido, SNPs) y una sola medición de una región cerebral. A pesar del gran éxito de los enfoques univariados, dada la perspectiva actual de los estudios IG en los que se analizan relaciones de todo el genoma frente toda la información del cerebro, el desarrollo de nuevos métodos estadísticos deviene crucial.

El objetivo principal de esta tesis consiste en investigar los determinantes genéticos relacionados con los cambios estructurales del cerebro, que a su vez afectan a los dominios del neurodesarrollo. Proponemos la aplicación y desarrollo de estrategias estadísticas para mejorar la evaluación de las relaciones biológicas asociadas a los dominios del neurodesarrollo. Específicamente, centramos nuestros esfuerzos de investigación en entender qué cambios genéticos en la estructura cerebral permiten mejorar la identificación de factores de riesgo associados con los dominios del transtorno por déficit de atención e hiperactividad y procesos cognitivos relacionados, como la función de atención.

PREFACE

This thesis was written at the Barcelona Institute for Global Health (ISGlobal) between October 2014 and January 2018, and it was supervised by Prof. Jordi Sunyer and Dr. Juan Ramón González.

The thesis consists of a compilation of 9 articles (5 published, 2 under review, 2 in preparation). The scientific publications co-authored by the Ph.D candidate are in agreement with the procedures of the Biomedicine Ph.D program of the Department of Experimental and Health Sciences of Universitat Pompeu Fabra. All publications based on data from the BREATHE (BRain dEvelopment and Air polluTion ultrafine particles in scHool childrEn) project in Barcelona, the population-based INMA- "INfancia and Medio Ambiente" Birth Cohort Project in Spain and the Rotterdam Study in Netherlands.

The present work contributed to 1) increase the knowledge about statistical modelling in the context of Imaging Genetic studies, 2) the standardization of analytical methods not-well established, 3) the development of new methods to increase statistical power in this field, 4) the characterization of the underlying biology of Attention/Deficit-Hyperactivity disorder symptoms and related neurodevelopmental domains, such as executive function and attention function, and 5) the offeriment of suggestions and recommendations for future research in Imaging Genetics.

Apart from the original papers included in the present thesis, in which the PhD candidate was responsible for all the statistical analyses and writing of the articles, she also co-authored 13 papers. She has been the main statistical analyst of the genetic studies in the BREATHE-European Research Council project, which in addition to the statistical analysis

and interpretation, also included the quality control and imputation of the genotyped data of the project, the quality control and extraction of brain structure data (surface-based measures) and the preparation of protocols [see *Annex*].
She has also participated/collaborated as the main analyst of the genetic studies in the EGG (Early Growth Genetics) Consortium and EAGLE (The EArly Genetics and Lifecourse Epidemiology) Consortium. She has established a position of assistant lecturer in the subject of Bioinformatics in the Master in Omics Data Analysis from the University of Vic-Universitat Central de Catalunya. She has been the co-organizer of the first and second Symposium for students from the Spanish Biometric Society and she has been the co-organizer of the second family meeting of INMA project, for which she also developed fieldwork and laboratory tasks.

RATIONALE

The study of brain development has increased dramatically over the last two decades, specifically in children. In this context, neuroimaging data and sample sizes are increasing which is improving the chances of using these data to understand neurodevelopment, cognition, and behaviour at the brain level, as well as to understand how these processes can be affected by environmental factors. Since the appearance of the International Human Genome Project (a catalogue of human genetic variants and their underlying correlations) and the ENCODE Consortium (Encyclopedia of DNA Elements) among others, we are starting to know much about the role of our genetic information. For instance, the interpretation and understanding of brain structure and functioning can be more accurate if genomic information is also considered. The joint analysis of genetic and neuroimaging data, a field known as Imaging Genetics (IG), provides a powerful tool to explore biological mechanisms which can affect the development of complex neurocognitive diseases.

Attention-Deficit/Hyperactivity Disorder (ADHD) represents a formidable challenge for psychiatry and neuroscience because of its high prevalence, lifelong nature, complexity and substantial heterogeneity. Neuroimaging and genetic studies in ADHD have reinforced the consideration of diagnostics as a neurobiological entity. However, the genetic architecture of ADHD is highly complex and polygenic, and the genetic basis of structural and functional brain changes linked to this disorder is still not well established. Specifically, the integration of neuroimaging and genetics provide a strong strategy to understand the underlying neurobiological

mechanisms operating in the development of ADHD symptoms, such as executive and attention function. However, the high dimensionality and the specific nature of these data still represent a challenge in IG research. A suitable understanding of how changes in our DNA are related to the differences in our brain structure and how these differences affect neurodevelopmental domains requires new statistical methods.

OBJECTIVE

The aim of this thesis is to apply and develop new methodological strategies to investigate the role of genetic determinants and structural brain changes in neurodevelopmental domains.

The specific objectives are:

1. To elaborate a conceptual framework of statistical methods to identify opportunities and improve analytical strategies in imaging genetic studies focused on neurodevelopmental domains.
2. To examine the effect of genetic mechanisms underlying neurodevelopmental domains:
    a. To summarize the evidence on the key genetic and brain regions associated with Attention-Deficit/Hyperactivity Disorder (ADHD) symptoms and related neurodevelopmental domains, and the neuroimaging techniques used.
    b. To study genetic variations and brain structures associated with ADHD symptoms and attention function domains.
    c. To study whether genetic risk variants for ADHD predict longitudinal reduction of brain structures, which in turn are related with other neurodevelopmental domains.
3. To develop analytical methods to improve results in Imaging Genetics studies:

a. To evaluate the best statistical model to fit quantitative neurodevelopmental scores based on a count of symptoms.

b. To develop gene-set analysis algorithms to test multiple Single Nucleotide Polymorphisms (SNPs) in a gene in relation to neurodevelopmental domains.

c. To develop a novel multivariate matrices integrative method to analyze genetic and neuroimaging data in relation to neurodevelopmental domains.

BASIS OF THIS THESIS

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Pujol J, Sunyer J, González JR. *Strategies for integrative analysis in Imaging Genetic studies.* (*Under review, since 6th December, 2017. Neuroscience & Biobehavioral Reviews journal*) **[Chapter 2]**

Vilor-Tejedor N, Cáceres A, Pujol J, Sunyer J, González JR. *Imaging genetics in attention-deficit/hyperactivity disorder and related neurodevelopmental domains: state of the art.* Brain Imaging Behav. 2016 Dec 15. [Epub ahead of print] Review. PubMed PMID: 27981420. **[Chapter 3]**

Alemany S, **Vilor-Tejedor N**, Bustamante M, Pujol J, Macià D, Martínez-Vilavella G, Fenoll R, Alvárez-Pedrerol M, Forns J, Júlvez J, Suades-González E, Llop S, Rebagliato M, Sunyer J. A *Genome-Wide Association Study of Attention Function in a Population-Based Sample of Children.* PLoS One. 2016 Sep 22;11(9):e0163048. doi: 10.1371/journal.pone.0163048. PubMed PMID: 27656889; PubMed Central PMCID: PMC5033492. **[Chapter 4 Section 1]**

Vilor-Tejedor N, Alemany S, Forns J, Cáceres A, Murcia M, Macià D, Pujol J, Sunyer J, González JR. *Assessment of Susceptibility Risk Factors for ADHD in Imaging Genetic Studies.* J Atten Disord. 2016 Aug 17. pii: 1087054716664408. [Epub ahead of print] PubMed PMID: 27535943. **[Chapter 4 Section 2]**

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Mortamais M, Pujol J, Sunyer J, González JR. *Sparse multifactorial analysis reveals the role of cerebellar tissue volumes and molecular processes in ADHD dimensions. (Under revie,w since 5th October, 2017. International Journal of Methods in Psychiatric Research).* **[Chapter 4 Section 3]**

Vilor-Tejedor N and Calle ML. *Global adaptive rank truncated product method for gene-set analysis in association studies*. Biom J. 2014 Sep;56(5):901-11. doi: 10.1002/bimj.201300192. PubMed PMID: 25082012. **[Chapter 5 Section 1]**

Vilor-Tejedor N, Gonzalez JR, Calle ML. *Efficient and powerful method for combining p-values in Genome-wide Association Studies*. IEEE/ACM Trans Comput Biol Bioinform. 2015 Dec 22. doi: 10.1109/TCBB.2015.2509977. [Epub ahead of print]. PubMed PMID: 28055892**. [Chapter 5 Section 2]**

Vilor-Tejedor N, Ikram MA, Roshchupkin GV, Cáceres A, Alemany S, Vernooij MW, Niessen WJ, van Duijn CM, Bustamante M, Pujol J, Sunyer J, Adams HH, González JR. *Independent Multifactorial Analysis to analyze multiblock data in Imaging Genetic studies; Application to Executive Function* (in preparation). **[Chapter 5 Section 3]**

Vilor-Tejedor N, Ikram MA, Roshchupkin GV, Niessen WJ, Alemany S, Adams HH. *Genome-wide genetic risk variants for ADHD predict longitudinal changes of ventricle structures in a population-based sample* (in preparation). **[Chapter 6]**

TABLE OF CONTENTS

Chapter 1

General Introduction

1.1 Genetic basis of complex traits

"Natural selection is a mechanism for generating an exceedingly high degree of improbability" - Ronald A. Fisher

Improbability and randomness occur with regularity. For instance, these occur in genetic variation, in the occurrence of mutations, in interactions between genes and the environment, and in the formation of neural networks during brain development. But even with the high degree of randomness, the growing evolution in the field of genetics has allowed genetic variation to be described at an unprecedented level, pinpointing single nucleotide changes that directly or indirectly affect a phenotype of interest, or a complex disease.
Genetic contribution to disease risk has been inferred in aggregation studies that quantify the increased disease risk in relatives of affected individuals. Unlike Mendelian diseases that are mainly governed by a single genetic mutation, common human diseases have a complex etiology: they are caused by the combined effects of environmental and multiple genetic factors.
Deoxyribonucleic acid (DNA) contains the genetic information used in the development and functioning of organisms. The human DNA sequence is composed of four chemical elements (nucleotides); Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) and is very similar among individuals ($> 99\%$). However, on average 1 in every 100-300 base pairs (pairs of nucleotides) differs between unrelated individuals in a population. There are different types of genetic variations present in the DNA sequence. The most common type of individual genetic variation corresponds to

single base substitutions which are termed single nucleotide polymorphisms (SNP). When a genetic variant is classified as an SNP, at least 1% of the population does not carry the same nucleotide at a specific position in the DNA sequence. Throughout the human genome there are about 10 million SNPs. Genetic variation at the SNP level is supposed to contribute to individual differences in susceptibility to complex traits and diseases, and to responses to treatment.

Figure 1. A) Genotype definition from a precise position along the DNA sequence. B) Inheritance genetic model categorization. C) Design matrix considering the additive genetic model. Each row represents the information of an individual. Each column represents the codification of an SNP for each individual.

These individual genetic differences contribute to phenotypic variation among individuals. Humans have two copies of

every sequence of DNA, one from each parent. Hence, most SNPs have two possible nucleotide variations (alleles); the most frequent one in the population is denoted by "A" and the less frequent one by "a". Since the human genome is diploid, that is, the DNA is duplicated in each cell of an individual, this yields three possible genotypes per SNP: "AA" for the common homozygous subjects, "Aa" for the heterozygous subjects and "aa" for the variant homozygous subjects [Figure 1a]. Standard inheritance genetic models include additive, recessive and dominant models. Assuming that carrying allele "a" is associated with an increased risk of phenotype/disease; an additive genetic model indicates that the susceptibility to develop the phenotype/disease increases linearly depending on the number of risk alleles, that is zero for "AA", one for "Aa" and two for "aa". Recessive genetic models indicate that two copies of risk allele are required for an increase in phenotype/disease risk, that is zero for "AA/Aa", and one for "aa". Dominant genetic model indicates that either one or two copies of the risk allele are required for an increase in phenotype/disease risk, that is one for "Aa/aa", and zero for "AA" [Figure 1b]. From a statistical point of view, an SNP can be thought of as a categorical variable with three different categories in the population that can be recoded numerically, assuming a prespecified genetic model [Figure 1c]. The study of the genetic basis of complex traits and diseases has changed dramatically in the last few decades.

The field of complex genetic diseases had been essentially constrained to candidate genetic studies. However, this design has been largely replaced by a more powerful hypothesis-free design: genome-wide association analysis (GWAS). GWAS consists of conducting genetic association analysis focused on associations between SNPs and phenotypes or disease status.

This approach usually uses a case-control study design in which the frequency of the minor allele of an SNP is compared between cases and controls. Nowadays, the number of genotyped SNPs for each individual is of the order of hundreds in candidate gene studies and one million or more in GWAS [Figure 2a].

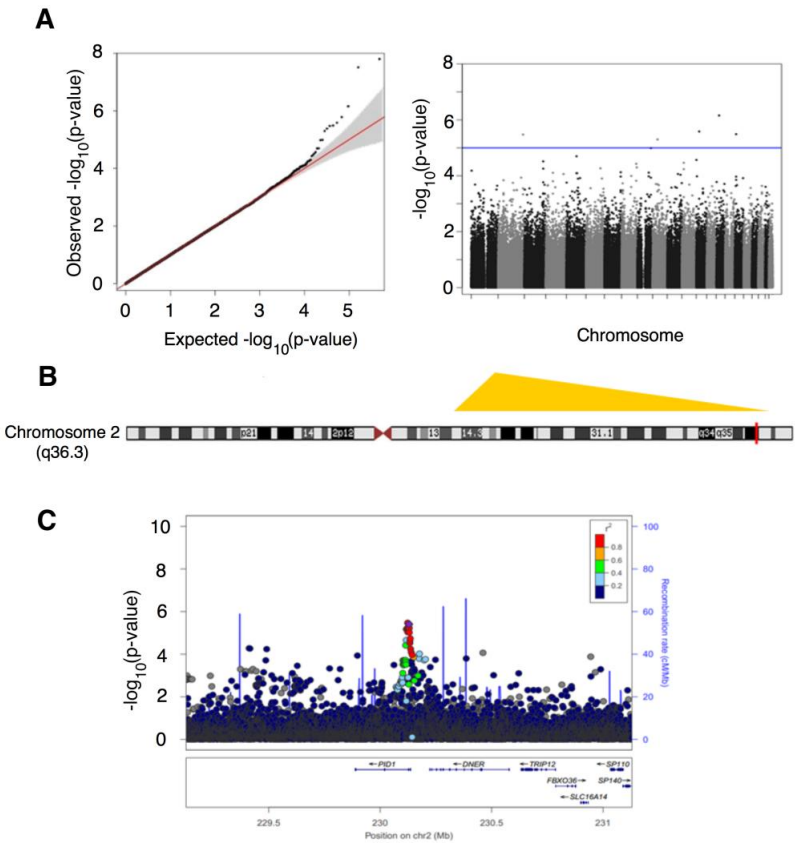

Figure 2. A) Quantile-quantile and Manhattan plot of genome-wide association results. The blue line indicates the suggestive level of statistical significance (pvalue<1E-5). B) Diagram of position of a candidate genetic variant within the chromosome. The red line indicates the position. C) Regional association plot. The linkage disequilibrium (LD; r2) between the SNP in focus and its SNPs genotyped or imputed within 1Mb is showed in red (high LD) to blue (low LD). The recombination rate is plotted in blue according to HapMap (CEU) (from Alemany et al. 2016).

Instead of genotyping SNPs in a number of genes that are thought to have some relation with the disease, as in candidate gene studies, GWAS are designed to cover most of the human genetic variation by genotyping SNPs across the whole genome without any prior hypothesis of causality. Indeed, GWAS are indirect association studies where the genotyped SNPs act as markers of nearby regions [Figure 2b]; it is assumed that an associated SNP will be either a causing disease variant or will be in linkage disequilibrium (LD) with an unmeasured causing variant [Figure 2c].

In the last few years, several GWAS have identified numerous genetic loci associated with disease risk or susceptibility for traits such as allergic eczema, birth weight or body mass index, many of which have been identified for the first time[2–4]. However, these well established variants only explain a small proportion of the phenotypic variance[5] and a small proportion of phenotypic variance due to additive genetic variability[6].

1.2 Neuroimaging techniques for brain exploration

Neuroimaging is based on various techniques to either directly or indirectly explore the structure or function of the brain. The two broad categories of neuroimaging are structural imaging, which deals with the static physical characteristics of the brain, and functional imaging, which looks at the brain in action, revealing dynamic changes in brain physiology. In this context, the most common neuroimaging tool to study the neurobiological substrates underlying the cognitive performance in humans is the magnetic resonance imaging (MRI) technique. MRI is used to construct a three-dimensional image of the layout of the brain. The MR signal is created by applying a strong magnetic field across the

brain. The single protons that are found in water molecules in the brain have weak magnetic fields. These fields are oriented randomly, but when the strong external field is applied a small fraction of them will align themselves with the field. The external field is applied constantly during the scanning process and when the protons are in the aligned state a brief radio frequency pulse is applied that knocks the orientation of the aligned protons. This new state produce a detectable change in the magnetic field and this is what forms the basis of the MR signal. Different types of images can be created from different MR signals.

The most frequently used method of structural analysis is voxel-based morphometry (VBM). This technique produces a measurement which offers a better discrimination between white matter and grey matter concentration[7]. VBM uses spatial co-registration to normalize individual brains into the coordinates of a brain template, allowing one to calculate either the density or volume of grey and white matter at each point in the brain. In addition, VBM is widely used to correlate local differences in grey and white matter volume to individual differences in cognitive function, which may in turn be linked back to individual genetic differences[8]. MRI techniques not only allow the study of brain structure but also the study of brain function. MRI can be adapted for use in detecting the changes in blood oxygenation associated with neural activity. This strategy is called functional Magnetic Resonance Imaging (fMRI). The basis for fMRI concerns the fact that neuronal activity in the brain is accompanied by an increased consumption of glucose and oxygen. In addition, there are pronounced changes in blood supply to the activated regions, characterized by increased cerebral blood flow (CBF) and cerebral blood volume (CBV)[9]. The MR signal used in fMRI is sensitive to the amount of

deoxyhaemoglobin in the blood. When neurons consume oxygen they convert oxyhaemoglobin to deoxyhaemoglobin, which in turn, has strong properties such as the introduction of distortions in the local magnetic field. This distortion is measured by the concentration of deoxyhaemoglobin present in the blood (BOLD signal) and is used to evaluate the increase in neural activity over time[10] (haemodynamic response function; HRF). Since the human brain is always physiologically active, functional imaging needs to measure relative changes in physiological activity. The most basic experimental design consists of subtracting the activity in each part of the brain whilst doing one task from the activity when doing a slightly different task (usually the activity is compared with a resting state). Neuroimaging is commonly used as an intermediate phenotype between the genetic information and the neurobehavioral development to elucidate genetic mechanisms of brain structure and function affecting neurodevelopmental domains[11].

1.2.1 Intermediate phenotype measures

Genetics play an important role and provided valuable insights into the genetic architecture of many neurodevelopmental domains and neurological diseases. Since now, genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits. Focus on neurodevelopmental domains and complex neurological diseases, the number of genetic variables identified is considerably reduced. This raises the question of using neuroimaging-based features as intermediate phenotypes, allowing a biological more plausible manner to assess these complex associations. An intermediate phenotype (often referred to as an endophenotype) is a quantitative

biological trait that is reliable and reasonably heritable[12]. Finding relationships between a primary outcome (disease status) and genetic factors is generally the main focus of genetic association studies.

However, in some contexts, clinical diagnosis is difficult to assess and provides inconclusive measures, due to the existence of heterogeneity between individuals with the same diagnosis and/or the comorbidity with other diseases. In addition, most neurodevelopmental domains and/or behavioural diseases are very complex at the level of genetic architecture/genetic inheritance (i.e. polygenic).

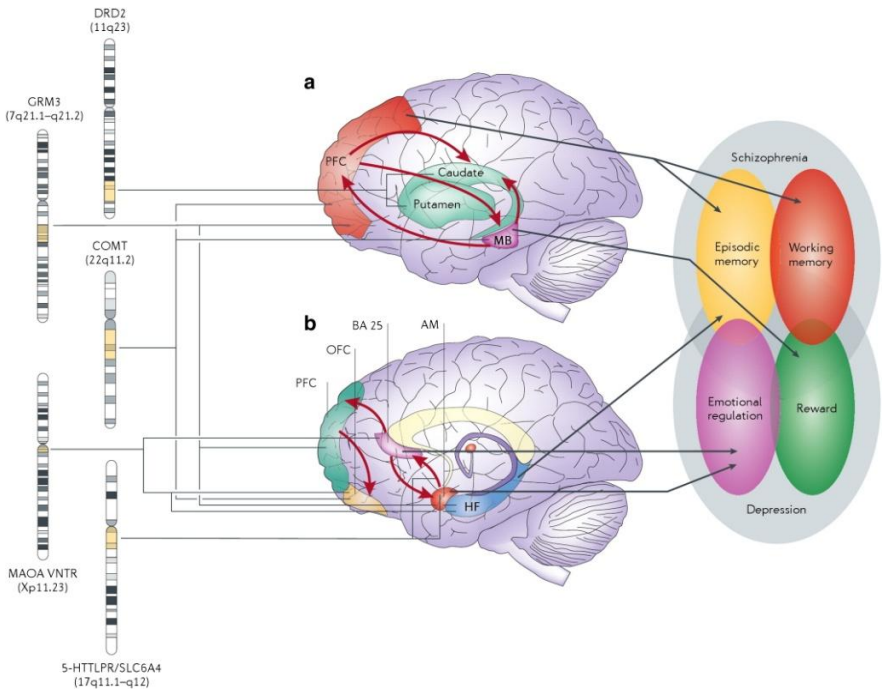

Figure 4. The complex path from genes to behaviour and disease phenotype: mediation through brain circuitry. Multiple genetic risk variants affect, through interaction with each other and the environment, multiple neural systems linked to several neuropsychological and behavioural domains that are impaired, in differing proportions, in psychiatric diseases. Abbreviations: BA 25 = Brodmann's area 25, HF = hippocampal formation, OFC = orbitofrontal cortex. (from Meyer-Lindenberg & Weinberger, 2006).

The intermediate phenotype strategy provides some advantages over that of a disease status, both in improving the power for discovering genetic risk variants, and in understanding underlying mechanisms of neurodevelopmental domains. For instance, the association of a particular behavioural feature with thickness in a specific cortical area may be a clue that the mechanisms involved in the development of that cortical region are also relevant to that behaviour [Figure 4]. Moreover, intermediate phenotypes are supposed to be more "simple" at a genetic level. Genetic effects on these measures would be stronger and independent of clinical status, thus, it is possible to detect them in the general population samples and genetic associations could be captured more objectively and accurately[11,13–17].

The intermediate phenotype strategy has lead to the collection of hundreds to thousands of neuroimaging and neuropsychological phenotypes to be used as intermediate phenotypes. However, it must be borne in mind that the availability of intermediate phenotypes supposes an increase in the amount of data for analysis. Therefore, developing statistical methods to incorporate multiple phenotypes in the same model becomes essential.

1.3 Neurodevelopmental domains

Neurodevelopmental domains are understood as any trait related to cognition, behaviour or brain function/structure, measured during childhood or adolescence. Recent evidence suggests that neurodevelopmental domains may exist as an extreme of a quantitative measure[18,19].

So much so that this framework of continuous symptom levels is already considered in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)[20] and dimensional

approaches are evaluated as a part of the National Institutes of Mental Health's Research Domain Criteria[21]. However, most studies, including genetic association studies, usually compare variation across genotypes based on diagnosis categories. Although the use of a categorical diagnosis can simplify the interpretability, it may result in biased estimates compared with quantitative measures[22]. Furthermore, classification of neurodevelopmental domains does not capture the variability in the population since, for instance, individuals who present extreme score values are assigned in the same group as those individuals with symptoms just above the diagnosis threshold. Moreover, the categorical diagnosis contrasts with classifying symptoms along a continuum, or a dimensional spectrum from normal to dysfunctional[23].

For research purposes, the use of a continuous score provides more power and allows the application of more advanced statistical methods[24,25]. Several studies have shown a better comprehension of the continuum of neurodevelopmental domains and a better etiological understanding of the development of symptoms related with attention and hyperactivity problems in the general population[26–28]. It would be beneficial, however, to study behaviour problems such as executive function, attention function, inattention and hyperactivity symptoms/traits, in order to gain a better understanding of related developmental disorders like Attention/Deficit-Hyperactivity disorder (ADHD).

1.3.1 Attention-Deficit/Hyperactivity Disorder symptoms

ADHD is a prevalent childhood neurodevelopmental disorder characterized by inappropriate symptoms of hyperactivity, impulsivity and inattention with an estimated worldwide

prevalence of 5.2%[29–31]. Heritability estimates from twin studies show a strong genetic component ranging from 60 to 90%, and sibling relative risk estimates range from a four- to an eight-fold increase[32,33]. Hence, ADHD is a complex disorder in which both environmental and genetic factors are implicated[34,35]. Among common studies, candidate genetic analyses have been carried out focusing mainly on genes involved in the dopaminergic and serotonergic pathways[36]. The most studied genes are the dopamine transporter gene (*DAT1*), the dopamine D4 (*DRD4*) and D5 (*DRD5*) receptor genes[37–39]. Most of the studies have yielded a number of replicated findings but meta-analyses showed that the associated variants exhibit small effect sizes, with odds ratios ranging from 1.1 to 1.9. The most promising results were reported in a recent GWAS of ADHD performed by the ADHD working group of the Psychiatric Genomics Consortium (PGC) and the Early Genetics and Lifecourse Epidemiology (EAGLE) Consortium[40]. The study, which included 20,183 ADHD cases and 35,191 controls, identified for the first time, 12 genome-wide significant independent loci (Pvalue < 1E-8). This GWAS, and the majority of genetic studies focused on ADHD, typically use a categorical status that indicates whether the disorder is present or absent[41–45]. Nevertheless, case-control designs, which have been applied in most GWAS, may penalize the statistical modelling of these symptoms, and in turn, the discovery of genetic risk factors associated with ADHD symptoms.

Neuroanatomical findings reported for ADHD are abnormal gray matter volume reductions, reduced volume of the basal ganglia, and cerebellum[46–48]. For the later, reduced cortical thickness of the frontal and temporal lobes and occipital areas have been reported, most being inconsistent[49–51].

The assessment of ADHD symptoms represents a formidable challenge for psychiatry, neuroscience and biology because of their high prevalence, lifelong nature, complexity, substantial heterogeneity, polygenic nature, and undefined biological characterization[52,53]. Moreover, the effects of ADHD genetic risk factors are likely to be mediated through impacts on brain structure and functioning. All of these issues motivate the emerging combination of neuroimaging-based features and genetic data. The integration of neuroimaging and genetics could provide a strong strategy to understand the underlying neurobiological mechanisms operating in the development of ADHD symptoms[54,55].

1.3.2 Attention function

Attention is a cognitive function essential in daily life. It has been proposed that three functionally and anatomically different networks are involved in this process: alerting, orienting and executive attention[56]. Alerting is defined as achieving and maintaining a state of high sensitivity to incoming stimuli; orienting is the selection of information from sensory input; and executive attention involves mechanisms for monitoring and resolving conflict among thoughts, feelings, and responses. The ability to measure changes in attention function may be very helpful both in considering disorders of attention, such as ADHD, schizophrenia, and Autism Disorder[57–59]. Thus, further research in attention function may have etiological implications for these disorders. For instance, we can expect genetic and brain structural influences on attention function development, which in turn, could be related to ADHD, and other complex neurodisorders.

1.4 Integrating genetics with neuroimaging

Joint analysis of genetic data and neuroimaging-based features, known as Imaging Genetics (IG), provides a better understanding of mechanisms of neurodevelopmental domains[60].

Until recently, IG studies started analyzing candidate imaging regions with candidate SNP/genes by using traditional correlation or linear regression models[61,62]. Later, IG studies focused on conducting GWAS to independently evaluate high-throughput SNP data with large-scale brain quantitative traits (QT)[63,64]. Nowadays, most studies in the field have begun to focus on performing simultaneously brain-wide genome-wide association analysis, where potential relationships are searched for in the whole genome and in the entire brain[65,66]. These studies use traditional multiple linear regression for association testing, allowing one to identify multilocus effects instead of individual genetic effects. However, these models perform extensive paired-wise independent tests between genetic and neuroimaging data sets.

Analogously to the genetic field, neuroimaging techniques produce a large amount of data. On average, each brain scan produces millions of 3D data points (voxels) containing information about a specific location in the brain. Therefore, just as it is difficult to select relevant SNPs and genes, it is also difficult to choose a subset of brain regions for analytical purposes.

Hence, the combination of genetic and neuroimaging fields becomes an analytical *big data* problem, because of the extremely large data sets to be analyzed. This produces several limitations such as: (i) multiple comparison correction which may lead to too strict significance thresholds and false

positives, (ii) the requirement of well-powered large studies which are not common in this field, and (iii) the infeasibility of most computational procedures. In this context, multivariate methods have gained attention for exploring relationships between multiple SNP and multiple QT[67–69].
Most of these methods reduce the dimensionality of the data for an easy interpretation of the results. However, multivariate association analysis methods are not easily applied. Despite the efforts thus far, novel methods have to be developed and standardized to deal with the superdimensionality and interpretability of the data, ideally, making use of the already available results from collaborative efforts to limit the analytical burden on human and computational resources.

1.5 Organization of the dissertation

The studies described in this dissertation are unified in advancing new directions and analytical perspectives in the field of imaging genetics. The innovative nature of both genetics and neuroimaging data, and even more importantly, their combination, has resulted in methodological and computational demands beyond current capabilities. The present work is organized as follows:
In **Chapter 2**, we review the current state-of-the-art of analytical strategies used in Imaging Genetics studies and we discuss the advantages of using multivariate statistics for this field.
In **Chapter 3**, we review the current knowledge of Imaging Genetics studies focused on ADHD domains and related neurodevelopmental domains. Main discoveries in genetics and neuroimaging fields are reported as well as the limited

use of multivariate statistical techniques to search for potential risk factors.

In **Chapter 4**, we explore the underlying neurobiological mechanisms operating in the development of ADHD and related neurodevelopmental domains from a two-step univariate perspective. In **Chapter 4 Section 1**, we aimed to identify SNPs associated with attention function. In the first analytical step we identified common genetic variants associated with attention function at a suggestive genome-wide level by performing a GWAS analysis. In the second step, we examined potential associations between relevant identified SNPs and variations in brain structure and function using neuroimaging tools. In **Chapter 4 Section 2**, we proposed and validated a new statistical modelization based on a negative binomial distribution to improve the modelling of ADHD symptoms in GWAS analysis. We applied the proposed statistical model in a GWAS analysis to select potential genetic features associated with ADHD symptoms in a first analytical step. Next, we examined potential associations between the identified significant SNPs and variations in brain structure. In **Chapter 4 Section 3**, we proposed the application of a multivariate cross-sectional framework based on a combination of a feature selection method and a multifactorial analysis in order to identify complex meaningful biological signals related to ADHD symptoms and Inattention/Hyperactivity domains.

Chapter 5 is focused on the proposal of novel multivariate methods for combining genetic data, neuroimaging features and/or neurodevelopmental domains. The first two methods proposed (**Chapter 5 Section 1 and Section 2**) utilize a gene-based test of association, which uses a permutational procedure and a semi-parametrical model, respectively, to derive gene-level p-values. Both methods were compared via

simulation analyses with GWAS results and conventional gene-set permutational procedures, showing a better performance in terms of statistical power and computational efficiency. Both proposed methods were included in the *globalGSA* R package and uploaded to the Comprehensive R Archive Network (CRAN) (See *Annex C*).

The third proposed method (**Chapter 5 Section 3**) is based on a multifactorial algorithm, which uses Independent Component Decomposition to derive relevant features from genetic and neuroimaging data, including a predictive step based on a multivariate regression. Additionally, the proposed method was compared via simulation analyses with the traditional Multifactorial Analysis and with univariate linear regressions. This method, as well as the previously proposed methods, are open-source and available via the GitHub repository[70].

In **Chapter 6**, in line with current research which has suggested that the effects of shared genetic liability may have longitudinal effects on brain structure, we explored whether genetic risk variants at the genome-wide level for ADHD predict longitudinal reduction of brain subcortical structures.

In **Chapter 7**, the main findings of all studies, contributions to the current knowledge, methodological considerations, strengths and limitations, and implications for future research are discussed.

Finally in **Chapter 8**, a general conclusion is presented.

Chapter 2

Strategies for integrative analysis in Imaging Genetics

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Pujol J, Sunyer J, et al. Strategies for integrated analysis in imaging genetics studies. Neurosci Biobehav Rev. 2018 Oct;93:57–70. DOI: 10.1016/ j.neubiorev.2018.06.013

Chapter 3

Imaging genetics in ADHD symptoms and related neurodevelopmental domains: state of the art

Vilor-Tejedor N, Cáceres A, Pujol J, Sunyer J, González JR. Imaging genetics in attention-deficit/hyperactivity disorder and related neurodevelopmental domains: state of the art. Brain Imaging Behav. 2017 Dec 15;11(6):1922–31. DOI: 10.1007/s11682-016-9663-x

Chapter 4

Understanding the underlying biological characterization of ADHD and related neurodevelopmental domains

Alemany S, Vilor-Tejedor N, Bustamante M, Pujol J, Macià D, Martínez-Vilavella G, et al. A Genome-Wide Association Study of Attention Function in a Population-Based Sample of Children. PLoS One. 2016 Sep 22;11(9):e0163048. DOI: 10.1371/journal.pone.0163048

Vilor-Tejedor N, Alemany S, Forns J, Cáceres A, Murcia M, Macià D, et al. Assessment of Susceptibility Risk Factors for ADHD in Imaging Genetic Studies. J Atten Disord. 2016 Aug 17;108705471666440. DOI: 10.1177/1087054716664408

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Mortamais M, Pujol J, et al. Sparse multiple factor analysis to integrate genetic data, neuroimaging features, and attention-deficit/hyperactivity disorder domains. Int J Methods Psychiatr Res. 2018 Sep;27(3):e1738. DOI: 10.1002/mpr.1738

.

Chapter 5

Novel methods for combining genetic variants, neuroimaging-based markers and/or neurodevelopmental domains.

Vilor-Tejedor N, Calle ML. Global adaptive rank truncated product method for gene-set analysis in association studies. Biometrical J. 2014 Sep;56(5):901–11. DOI: 10.1002/bimj.201300192

Vilor-Tejedor N, Gonzalez JR, Calle ML. Efficient and Powerful Method for Combining P-Values in Genome-Wide Association Studies. IEEE/ACM Trans Comput Biol Bioinforma. 2016 Nov 1;13(6):1100–6. DOI: 10.1109/TCBB.2015.2509977

Independent multifactorial analysis to analyze multiblock data: application to Imaging Genetics

Authors: **Vilor-Tejedor N**, Ikram MA, Roshchupkin GV, Cáceres A, Alemany S, Vernooij MW, Niessen WJ, van Duijn CM, Bustamante M, Pujol J, Sunyer J, Adams HH, González JR.

(in preparation)

Independent Multifactorial Association Analysis to analyze multiblock data in Imaging Genetics.

Authors: Vilor-Tejedor* N (1-3); Ikram MA (4); Roshchupkin GV (5,7); Cáceres A (1-3); Alemany S (1-3); Vernooij MW (4,5); Niessen WJ (5, 6, 7); van Duijn CM (4); Bustamante M (1-3,8), Pujol J (9); Sunyer J (1-3,10); Adams** HH(4,5); González** JR (1-3).

**** co-last authors**

(1) Barcelona Research Institute for Global Health (ISGlobal), Barcelona, Spain.

(2) Universitat Pompeu Fabra (UPF), Barcelona, Spain.

(3) CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

(4) Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands

(5) Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands

(6) Department of Medical Informatics, Erasmus MC, Rotterdam, the Netherlands

(7) Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands.

(8) Centre for Genomic Regulation (CRG), Barcelona, Spain

(9) MRI Research Unit, Hospital del Mar, and Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM G21, Barcelona, Spain

(10) IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

*Correspondence to: Natàlia Vilor-Tejedor, Barcelona Institute for Global Health (ISGlobal) . C. Doctor Aiguader 88, 08003 Barcelona, Spain. E-mail: natalia.vilor@isglobal.org. ORCID: 0000-0003- 4935-6721

CONFLICT OF INTEREST

The authors declare no conflicts of interest and financial disclosures.

Abstract

The joint analysis of high-dimensional heterogeneous data represents a challenge in studying complex biological systems. From a statistical point of view, the analysis of data involving multiple sources can be formulated as a multiblock framework. Imaging genetics studies that are focused on the joint analysis of genomic and imaging-based features are examples of such analyses. In this paper, we introduce a novel multifactorial algorithm, referred as Independent Multifactor Association Analysis (ICA-MFA), which uses Independent Component decomposition to derive relevant features from block data that improve the amount of variability explained. In addition, this approach incorporates a feature selection based on independent component regression. We evaluate the performance of ICA-MFA against multifactorial analysis (MFA) and univariate analysis in a simulation study. We evaluated the impact of the novel framework in an imaging genetics study of 4057 individuals belonging to the population-based Rotterdam Study with available genetic and neuroimaging data as well as information about cognitive functioning. Specifically, we used ICA-MFA to detect genetic features related to structural brain regions, which are known to play an important role in the mechanisms of executive function. We showed how ICA-MFA explains up to 10-fold more variance than the more common MFA and univariate methods.

1. Introduction

Current biomedical research increasingly combines high-throughput data. For instance, sequencing technologies produces omics data at different levels of cellular components. In addition, magnetic resonance imaging produces vast amounts of data (eg. Structural, functional and connectivity) with even more complex features and broader dimensions (Luo, Wu, Gopukumar, & Zhao, 2016). The analysis of this type of data presents different challenges, even more so if we consider the combined analysis of both sources, field referred as imaging genetics.

A common strategy to investigate potential associations between neuroimaging features and genetic data is based on performing massive marginal linear models in which extensive paired-wise correlations are computed (Hoogman et al., 2014). However, this strategy has important limitations such as (i) the inability to exploit the multidimensionality of data and synergistic effects between variables and (ii) the requirement of a large number of subjects for well-powered inferences.

Joint multivariate methods have the potential to better capture the complex relationships that may exist between different biological levels and significantly reduce the number of statistical tests, accounting for the multiple testing correction problems (Liu & Calhoun, 2014). Multiblock methods are an alternative to address problems regarding marginal analyses (Kawaguchi, Yamashita, & Alzheimer's Disease Neuroimaging Initiative, 2017).

Multiple Factor Analysis (MFA) is one of the most popular methods for analyzing multiple sets of variables measured on the same observations (Husson, Lê, & Pagès, 2011). MFA aims to provide common factor scores and measures of discrepancy between blocks of variables (Abdi, Williams, & Valentin, 2013). However, singular value decomposition in MFA is based on PCA, which is adequate only if the data is normally distributed, linear or stationary. Also, including strongly correlated variables

can overemphasize the contribution of the estimated principal components (Lever, Krzywinski, & Altman, 2017). To overcome these problems, we propose a novel method called Independent Multifactorial Analysis. This method is an extended implementation of MFA, where the component value decomposition is based on Independent Component Analysis that does not assume multivariate normality and linearity (Hyvärinen, 2013).

The main advantages of the proposed method are that (1) it is applicable if there is correlation between variables within structures, (2) it increases the variability explained by the data components, (3) it performs a feature selection considering the correlation data structure of variables.

This article is organized as follows. In section 2, we propose an extension of MFA, referred to as Independent Multifactorial Analysis (ICA-MFA). ICA-MFA incorporates Independent Component analysis (ICA) as a generalization of PCA decomposition. ICA-MFA also incorporates a feature selection based on a meaningful independent component regression (ICR). In section 3, we explore the performance of the algorithm. We evaluate and compare the performance of ICA-MFA with both the MFA method and traditional univariate analyses in a simulation study. In section 4, we applied and compare ICA-MFA, MFA and univariate analysis in an imaging genetics study using data from the Rotterdam Study (Ikram et al., 2017). The main results of simulations and real data analyses are discussed in the final section of the paper.

2. Method

Multiple Factorial Analysis (MFA) is a multivariate version of Factorial Analysis (FA) and an extension of PCA used to integrate m different sets of variables (in a matrix format), $X_1,...,X_m$ on the same set of observations. MFA is mainly comprised in three steps:

First, a PCA of each data set is performed via single value decomposition (SVD). The SVD of a given IxJ rectangular data matrix X is its factorization into three matrices

$$X = U\mathbb{\Gamma}V^T \text{ such that } U^TU = V^TV = I,$$

where U is a IxL matrix of the normalized left singular vectors, V is a JxL matrix of the normalized right singular vectors and $\mathbb{\Gamma}$ is the LxL diagnal matrix of the L singular values, L being the rank of the decomposed matrix X, and U and V being orthonormal matrices.

Second, data sets are normalized by dividing all the elements of each table X_i by the corresponding explained variance of the first singular vector, given by the inverse of the first squared singular value $\mathbb{\Gamma}_{i,(1,1)}$.

Finally, the normalized data sets are concatenated into a unique data set and a PCA is computed on the general data set to evaluate how much the whole set of variables contribute to the inertia extracted by a component.

To address problems related with the single value decomposition in PCA (orthogonality assumption and multivariate normal distribution of the variables in each dataset), we present a statistical methodology based on an extension of MFA, referred as Independent Multifactorial Analysis (ICA-MFA). This approach is designed to evaluate potential relationships between sources of data based on Independent Component Analysis (ICA) decomposition and Independent Component Regressions (ICR) that is used to link MFA results with an outcome of interest.

2. 1 Independent Component Analysis (ICA)

ICA aims to find a linear representation of non-Gaussian vectors such that the estimated vectors are statistically independent (Comon, 1994). ICA decomposition is similar to PCA model,

but while. The PCA identifies linear combinations of the original variables such that the covariance between the derived variables is zero. Whereas, ICA identifies variables that are statistically independent. As independence implies null covariance and not vice versa, it is a stronger condition that can better reflect the intrinsic properties of mixed signals.
The ICA decomposition of X is given by

$$X = AS,$$

where S and A are matrix of independent components and mixture matrix, respectively. Independent component regression (ICR) is similar to Principal Component regression (PCR), with the difference that ICR uses independent components S, and the coefficient matrix, A, obtained by ICA in the regression analysis instead of the principal components and matrix of scores obtained by PCA decomposition. Hence, since X can be described by its coefficient matrix, A, the multiple linear regression equation between A and the matrix of components, can be defined as in PCR (Bair, Hastie, Paul, & Tibshirani, 2006).

2. 3 Independent Multifactorial Analysis (ICA-MFA)

We propose a multiblock framework to evaluate relationships between two rectangular data matrices collected on the same set of observations. Although the method is described considering two data sets (genetic data and imaging features), the procedure can be extended to K matrices.
Consider an imaging genetics study where $N_{n.k}$ and $G_{n,p}$ denotes blocks of neuroimaging and genetic data, respectively, n stands for the number of individuals, k is the number of neuroimage-based features (i.e. brain volumes, …) and p denotes the number of genetic variants (i.e., SNPs, genetic scores, structural variants, …) the proposed algorithm comprises five steps:

Step 1. Computing ICA decomposition on each block of variables: An ICA decomposition of each block of variables is performed in order to search linear combinations of variables that optimize statistical independence. Let us assume *c* independent components $\Phi_1, \ldots, \Phi_c$. Therefore, by definition, the joint probability density function (pdf) is factorizable as the joint product of *c* terms.

Then, each dataset is decomposed into a linear mixture$x_1, \ldots, x_k$, of *k* independent components,

$$\begin{cases} x_{N_j} = a_{j_{1\Phi_{N_1}}} + a_{j_{2\Phi_{N_2}}} + \cdots + a_{j_{c\Phi_{N_k}}} \\ x_{G_z} = b_{z_{1\Phi_{G_1}}} + b_{z_{2\Phi_{G_2}}} + \cdots + b_{z_{c\Phi_{G_k}}} \end{cases}$$

where A_i, B_j are the associated mixing matrices with elements a_{jk}, b_{zk}.

$$\begin{cases} N_{nxk} = A \cdot S_N \\ G_{nxp} = B \cdot S_G \end{cases}$$

Step 2. Normalization of each data table: data sets are scaled by dividing all of its elements by the square root of the first independent component from those obtained in step 1, following the same strategy as in MFA.

$$\begin{cases} Z_1 = N_{nxk}\sqrt{\Phi_{N_1}^{-1}} \\ Z_2 = G_{nxp}\sqrt{\Phi_{G_1}^{-1}} \end{cases}$$

Step 3. Concatenation of data sets: The normalized datasets, $Z_{[1]}$ and $Z_{[2]}$ are concatenated into a complete dataset denoted by *C*.

$$C = \left[Z_{[1]}\right] \vee Z_{[2]}$$

Step 4. Compute an ICA on the generalized dataset J: ICA decomposition is performed on the concatenated table *J* to extract a vector of independent imaging genetic components Φ,

$$J = \Sigma \cdot \Phi_i$$

where Σ is the associated mixing matrix with elements σ_{yt}.

Step 5. Feature selection: Finally, an ICR is computed to determine relevant features related to our outcome of interest (dichotomous or quantitative trait), *Y*, and the total amount of variability explained for those features. Independent factors included on regression models are selected using the Akaike information criterion (AIC).

$$Y_j = \beta_0 + \sum \beta_j \Phi_{ij} + \varepsilon_i$$

3. Simulation Study

3.1 Simulation Design

We performed a simulation study to compare the variability explained by ICA-MFA, MFA, and univariate linear regression models for a quantitative trait. We use the *PhenotypeSimulator* package from GitHub (https://github.com/cran/PhenotypeSimulator).

PhenotypeSimulator functions fit a linear model with the genotype as the explanatory variable and the phenotype as the response variable, including the effect of additional covariates and random noise.

We simulated datasets with different sample sizes *n*, (n= 1000, 3000), a quantitative outcome *Y* mimicking disease score, n_s variables representing the genotypes of a set of Single Nucleotide Polymorphisms (n_s =*10, 100, 1000*), and n_b =15 image variables representing different brain structures. Genetic effects were simulated as the matrix product of genotype matrix, $n \times n_{sc} SNPs$, and effect size matrix $n_s \times Y$, assuming a Linkage Disequilibrium (LD) structure, an additive genetic model and allele frequencies of 5, 10, 30 and 40%. Allele frequencies were uniformly sampled and used to simulate individual genotypes by drawing values from a binomial distribution with 2 trials. Information was summarized using non-standardized allele codes (i.e; 0, 1, 2). Brain structure effects were simulated as

quantitative variables following a multivariate normal distribution. From realistic data (McCarthy et al., 2015; Table 2) we extracted the mean modulate values, μ_{n_b}, and standard deviations, σ_{n_b}, of 15 scanner-specific cortical thickness values for brain structures. Each source of data was scaled to explain a certain proportion of the entire outcome variance. We assumed that the proportion of variance explained by brain structure components was 30%, the total genetic variance 40%, and the proportion of variance of fixed genetic effects 2.5%. Moreover, single SNP effects were assumed of 1% of the total phenotypic variance. In total, we simulated 10 different scenarios assuming combinations of the considered parameters.
The information is summarized in Table 1.

3.2 Simulation Evaluation Performance

The performance of each method was compared by computing the variability of *Y* that was explained by ICA-MFA, MFA and univariate linear regressions, calculated as the explained sums of squares (ESS) due to regression. The ESS is defined as the sum of the squares of the differences of the predicted values and the mean value of the response value:

$$ESS = \sum_{i} (\hat{y}_i - \hat{y})^2$$

We additionally compared the goodness of fit for the different models based on an analysis of deviance with the inclusion of the latent components. We compared each model with models that do not include any component. Moreover, we tested whether the extracted principal and independent components were significantly associated with *Y*.

3.3 Simulation Results

Results of the simulation studies are summarized in Tables 2 to 5. Tables 2 and 3 provide the percentage of variability explained when varying the number of causal SNPs. We observed that in

all scenarios, ICA-MFA outperforms both MFA and univariate regression approach in terms of variability explained. Considering 3 components of ICA-MFA provide an increase of 30% of variance explained in all scenarios. We also showed that the magnitude of variability explained does not depend on the sample size or the number of imaging genetic covariates included in the models. However, as it was expected, the amount of variability explained highly depends on the number of causal SNPs included analysis.

Tables 4 and 5 provide the percentage of variability explained for different number of SNPs analyzed, in which only 10 SNPs were causal. As expected, the explained variability was lower than when all SNPs were causal. Again, ICA-MFA outperforms the MFA method and the univariate regression approach. Moreover, in all scenarios, independent components obtained from ICA-MFA provides a better goodness of fit based on the Akaike information criterion (AIC), and a better performance of prediction compared with components obtained from MFA [Figures 1-2]. For reproducibility purposes, the data sets and scripts supporting the results of these simulation studies can be found at *https://github.com/natvt8/ICA-MFA*.

4. Application to Real Dataset

We applied ICA-MFA and MFA methods on a subset of imaging genetics data from the Rotterdam Study (Ikram et al., 2017). We evaluate a set of 9 SNPs, and 39 subcortical surfaces for the same individuals, in relation to executive functioning. We obtained those features (brain structures, and genetic variations) that characterize the executive functioning, based on variability explained.

4.1 Study Population

The Rotterdam study is a prospective population-based cohort study comprising of 14,926 middle aged and elderly individuals,

investigating the determinants and consequences of age-related diseases in older adults (Ikram et al., 2017). Genotyping was performed on 11,496 individuals and 5,691 unique participants underwent brain magnetic resonance imaging (MRI).
From the total of 14,926 participants in the Rotterdam Study, genotypic, neuroimaging data and executive function were available for 4,057 individuals (mean age 64.7) [Figure S1]. The Rotterdam study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. A written informed consent was obtained from all participants.

4.2 Cognitive functioning measures

Executive cognitive function was assessed with the Letter-Digit Substitution test (LDST: Jolles et al. 2017). The LDST asks the participants to make as many letter-digit combinations as possible in 60 seconds, following an example that shows the correct combinations. Normative LDST have been well established for adults (van der Elst, van Boxtel, van Breukelen, & Jolles, 2006). In the Rotterdam Study all LDST were administered by trained investigators in quiet rooms. The test took no longer that 30 minutes to complete and a stopwatch was used for the control of time (Hoogendam, Hofman, van der Geest, van der Lugt, & Ikram, 2014). LDST were assessed from 1997-1999 and consecutive follow-up examinations every 3 to 4 years have been conducted until now. For analytical purposes, in this study, we selected those executive function measurements closest to the brain MRI performed in the study participants.

4.3 Image acquisition, processing and selection.

Magnetic Resonance Imaging scanning was done on a 1.5-T MRI scanner (Signa Excite II; General Electric Healthcare, Milwaukee, WI, USA). The MRI protocol included a high-

resolution axial T1-weighted 3-dimensional fast radio frequency spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse (FASTSPGR-IR) sequence (repetition time [TR] = 13.8 ms, echo time [TE] = 2.8 ms, inversion time [TI] = 400 ms, field of view [FOV] = 25 cm2, matrix = 416 × 256, flip angle = 20°, number of excitations [NEX] = 1, bandwidth [BW] = 12.50 kHz, 96 slices with slice thickness 1.6 mm 0-padded to 0.8 mm). All slices were contiguous. According to the Rotterdam Study standard acquisition protocol images were resampled to 512 x 152 x 192 voxels (voxel size: 0.5 x 0.5 x 0.8 mm3) (Ikram, et al., 2016).
The T1-weighted MRI scans were processed using a model-based automated procedure of Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) (Fischl et al., 2004) to obtain segmentations and volumetric summaries of subcortical structures and thickness of the cerebral cortex. This procedure automatically assigns a neuroanatomical label to each voxel in an MRI volume based on probabilistic information obtained from a manually labeled training set. This yielded intracranial volume (ICV) and gray and white matter volumes for cerebellum and cerebrum. Further details of the MRI protocol can be found in Ikram et al., 2017. For the purpose of this study, we selected 39 subcortical structures [Table S1].

4.4 Genotyping acquisition and genetic variant selection
The Rotterdam Study consist on three subcohorts, which were genotyped with the 550K (cohort 1), 550K duo (cohort 2) and 610K (cohort 3) Illumina arrays. Samples with a call rate below 97.5%, gender mismatch, excess autosomal heterozygosity (>0.336), duplicates or family relations and ethnic outliers were excluded. Genetic variants were filtered by Hardy-Weinberg equilibrium ($P<10^{-6}$), allele frequency (excluding minor allele frequency (MAF < 0.001) and SNP call rate with a minimum of 98%. Genotypes were imputed using MACH/minimac software to the 1000 Genomes phase I version 3 reference panel (all

populations). Among the variants imputed, a total of 9 loci associated with Attention-Deficit/Hyperactivity Disorder (ADHD) at a genome-wide threshold of significance ($P<10^{-8}$), were pre-selected [Table SM2] (Demontis et al., 2017). Moreover, we constructed a genetic risk score (GRS) by multiplying the number of risk alleles by their reported odds ratio (after natural logarithm transformation) for the disease, and summing this weighted allele score of each variant up into a disease risk score for ADHD.

4.5 Results

Variability explained by each component

ICA-MFA identified three independent components (Φ) that pass significance criteria Φ_1 (P=2.22E-94); Φ_2 (P=9.03E-08); Φ_3(P=2.71E-88), explaining approximately 18% of the global variance of executive function, while the first three principal components (PCs) from MFA were only able to detect 1% [Table 6]. Specifically, we show how the increment in the variability explained for executive function varies by only 1% when going from incorporating a main component to three main components in the MFA procedure. For ICA-MFA, the amount of variability explained is around 9% considering one IC, increases to 10% when considering two components, and reaches 18% when including all three components. It seems then that the first and third components would be the most representative in the quantification of the total variability of executive function.

Contribution of variables to each dimension

Figures from 2 to 4 show those variables contributing the most to the definition of the three dimensions of ICA-MFA. Variables that contribute the most to the first dimension are lateral ventricle volumes, cerebellar cortex, cerebellum, white matter, and hippocampus volumes. Variables that contribute to the

second dimension are white matter and gray matter volumes, and also cerebellar cortex. Finally, variables that contribute to the third dimension are gray matter, cerebellar cortex, lateral ventricles and corpus callosum volumes. For this third dimension we additionally appreciate the contribution of three genetic components, rs1427829 (*DUSP6/POC1B*), rs4858241 (*Intergenic*) and rs9677504 (*SPAG16*).

Compared to ICA-MFA, MFA selects a larger number of genetic variables in quantifying the relationship with the dimensions of the components, while the dimensions of ICA-MFA are mostly characterized by neuroimaging-based features. It is also observed how neuroimaging-based features selected by the MFA method correspond to general measures such as white matter (WM), gray matter (GM) and total brain volume (TBV) [Figures S2-S4]. ICA-MFA seems to better specify the structure-based features that determine the dimensional components. Moreover, none of the dimensions of ICA-MFA and MFA are characterized by the influence of the genetic risk score.

4. Discussion

The aim in the field of imaging genetics is to find relations between genetic data and imaging phenotypes using large datasets; these relations often have a small effect size (Abi-Dargham & Horga, 2016; Medland, Jahanshad, Neale, & Thompson, 2014). In order to increase statistical power, new methods and technologies for data reduction are being considered. We developed a new method that better explains variability in multifactorial analyses than conventional methods. Our method incorporates independent component decomposition instead of the more common principal component analysis. The proposed multifactorial method derives an integrated picture of the observations and the relationships between the groups of

variables. By taking advantage of the independent component decomposition, the proposed method outperforms multifactorial analysis and univariate regressions in a simulation study. Moreover, in a real life proof of principle study, the explained variability accounted for by our proposed method is higher, demonstrating the potential of the algorithm.

We explored the performance of the proposed algorithm on a subset of imaging genetics data from the population-based Rotterdam Study, in which we explored genetics and imaging features in relation to executive cognitive function in an adult population sample. Instead of independently performing univariate regressions, or applying MFA , we integrated the multimodal feature datasets applying independent component decomposition. While univariate results and multimodal MFA combinations only explained a limited proportion of variability (less than 2%), the proposed ICA-MFA increased the explained variability (%) and allowed the identification of significant independent components that maximize the variability explained. Though meant primarily as a proof of principle example, from a biological perspective, the results obtained in this real data sample provide new views of research on the characterization of the cognitive processes that underlie more complex symptoms such as ADHD (Curatolo, D'Agati, and Moavero 2010; Purper-Ouakil et al. 2011). The biological characterization of executive functioning could suggest an approximation of the joint affectation that genetic profiles and changes at the level of brain structure on these symptoms and related neurodevelopmental domains (Mueller & Tomblin, 2012; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005).

The potential application of the ICA-MFA algorithm on imaging genetics studies constitutes an important aspect of integrating imaging genetics data, especially in relation to neurodevelopment domains due to the small number of studies and inconsistency of the results (Vilor-Tejedor, Cáceres, Pujol,

Sunyer, & González, 2016). Hence, further research may greatly benefit from the development of multivariate approaches which represent a potential form to increase the statistical power to detect significant causal factors in multiblock data analysis (Meyer-Lindenberg, 2012).

ACKNOWLEDGEMENTS

Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2017 FI_B 00636), Generalitat de Catalunya – Fons Social Europeu. This work has been partially supported by a STSM Grant from EU COST Action 15120 Open Multiscale Systems Medicine (OpenMultiMed) and Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP). Further support was obtained through the Ministerio de Economía e Innovación (Spain), grant MTM2015-68140-R. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya. Silvia Alemany thanks the Institute of Health Carlos III for her Sara Borrell postdoctoral grant (CD14/00214).

The generation and management of GWAS genotype data for the Rotterdam Study are supported by the Netherlands Organization of Scientific Research NWO Investments (no. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project no. 050-060-810. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. This research is supported by the Dutch Technology Foundation STW (12723), which is part of the NWO, and which is partly funded by the Ministry of Economic Affairs. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project: ORACLE, grant agreement No: 678543).

REFERENCES

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(2), 149–179. https://doi.org/10.1002/wics.1246

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nature Medicine*, *22*(11), 1248–1255. https://doi.org/10.1038/nm.4190

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137. https://doi.org/10.1198/016214505000000628

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, *36*(3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9

Curatolo, P., D'Agati, E., & Moavero, R. (2010). The neurobiological basis of ADHD. *Italian Journal of Pediatrics*, *36*(1), 79. https://doi.org/10.1186/1824-7288-36-79

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., … Neale, B. M. (2017). Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv*, 145581. https://doi.org/10.1101/145581

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*, S69–S84. https://doi.org/10.1016/j.neuroimage.2004.07.016

Hoogendam, Y. Y., Hofman, A., van der Geest, J. N., van der Lugt, A., & Ikram, M. A. (2014). Patterns of cognitive function in aging: the Rotterdam Study. *European Journal of Epidemiology*, *29*(2), 133–140. https://doi.org/10.1007/s10654-014-9885-4

Hoogman, M., Guadalupe, T., Zwiers, M. P., Klarenbeek, P., Francks, C., & Fisher, S. E. (2014). Assessing the effects of common variation in the FOXP2 gene on human brain structure. *Frontiers in Human Neuroscience*, *8*, 473. https://doi.org/10.3389/fnhum.2014.00473

Husson, F., Lê, S., & Pagès, J. (2011). *Exploratory multivariate analysis by example using R*. CRC Press. Retrieved from https://www.crcpress.com/Exploratory-Multivariate-Analysis-by-Example-Using-R/Husson-Le-Pages/p/book/9781439835814

Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *371*(1984), 20110534. https://doi.org/10.1098/rsta.2011.0534

Ikram, M. A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., … Hofman, A. (2017). The Rotterdam Study: 2018 update on objectives, design and main results. *European Journal of Epidemiology*, *32*(9), 807–850. https://doi.org/10.1007/s10654-017-0321-4

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (n.d.). The Maastricht Aging Study: Determinants of cognitive aging. Retrieved from http://www.np.unimaas.nl/maas

Kawaguchi, A., Yamashita, F., & Alzheimer's Disease Neuroimaging Initiative. (2017). OUP accepted manuscript. *Biostatistics*, *18*(4), 651–665. https://doi.org/10.1093/biostatistics/kxx011

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, *8*, 29. https://doi.org/10.3389/fninf.2014.00029

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, *8*, 1–10. https://doi.org/10.4137/BII.S31559

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J.-A., Coman, I. L., & Kates, W. R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, *9*, 379. https://doi.org/10.3389/fnins.2015.00379

Medland, S. E., Jahanshad, N., Neale, B. M., & Thompson, P. M. (2014). Whole-genome analyses of whole-brain data: working within an expanded search space. *Nature Neuroscience*, *17*(6), 791–800. https://doi.org/10.1038/nn.3718

Meyer-Lindenberg, A. (2012). The future of fMRI and genetics

research. *NeuroImage*, *62*(2), 1286–1292. https://doi.org/10.1016/j.neuroimage.2011.10.063

Mueller, K. L., & Tomblin, J. B. (2012). Diagnosis of ADHD and its Behavioral, Neurologic and Genetic Roots. *Topics in Language Disorders*, *32*(3), 207–227. https://doi.org/10.1097/TLD.0b013e318261ffdd

PURPER-OUAKIL, D., RAMOZ, N., LEPAGNOL-BESTEL, A.-M., GORWOOD, P., & SIMONNEAU, M. (2011). Neurobiology of Attention Deficit/Hyperactivity Disorder. *Pediatric Research*, *69*(5 Part 2), 69R–76R. https://doi.org/10.1203/PDR.0b013e318212b40f

van der Elst, W., van Boxtel, M. P. J., van Breukelen, G. J. P., & Jolles, J. (2006). The Letter Digit Substitution Test: Normative Data for 1,858 Healthy Participants Aged 24–81 from the Maastricht Aging Study (MAAS): Influence of Age, Education, and Sex. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 998–1009. https://doi.org/10.1080/13803390591004428

Vilor-Tejedor, N., Cáceres, A., Pujol, J., Sunyer, J., & González, J. R. (2016). Imaging genetics in attention-deficit/hyperactivity disorder and related neurodevelopmental domains: state of the art. *Brain Imaging and Behavior*. https://doi.org/10.1007/s11682-016-9663-x

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the Executive Function Theory of Attention-Deficit/Hyperactivity Disorder: A Meta-Analytic Review. *Biological Psychiatry*, *57*(11), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

Scenario	N	NSNPs	NcSNPs
1	1000	10	10
2		100	100
3		1000	1000
4	3000	10	10
5		100	100
6		1000	1000
7	1000	100	10
8		1000	10
9	3000	100	10
10		1000	10

Table 1. Characteristics of the simulated scenarios. Legend: N, Number of individuals; NSNPs, Number of SNPs; NcSNPs, Number of causal SNPs; NROIs, Number of simulated brain regions (15); Proportion of fixed variance from brain structure components (0.3); h2, Proportion of variance of fixed genetic effects (0.2).

N=1000			
NSNPs=10, NcSNPs=10			
	ICA-MFA	**MFA**	**univariate**
c=1	2.0%	0.46%	0.05%
c=1,2	35.81%	0.50%	
c=1,2,3	40.83%	0.53%	
NSNPs=100, NcSNPs=100			
	ICA-MFA*	**MFA****	**univariate**
c=1	9.76%	0.90%	0.05%
c=1,2	21.28%	0.91%	
c=1,2,3	44.55%	1.59%	
NSNPs=1000, NcSNPs=1000			
	ICA-MFA*	**MFA****	**univariate**
c=1	0.91%	0.25%	0.05%
c=1,2	7.17%	1.03%	
c=1,2,3	33.01%	1.17%	

Table 2. Percentage of variability explained depending on the number of components, c, included in the model and the number of SNPs. Scenarios from 1 to 3. Legend: N, Number of individuals; NSNPs, Number of SNPs; NcSNPs, Number of causal SNPs; NROIs, Number of brain structures. ICA-MFA, Independent Multifactorial Association method MFA, Multifactorial Analysis.

N=3000 *NSNPs=10, NcSNPs=10*			
	ICA-MFA*	**MFA****	**univariate*****
c=1	15.94%	0.95%	
c=1,2	27.82%	1.48%	0.04%
c=1,2,3	44.08%	1.69%	
NSNPs=100, NcSNPs=100			
	ICA-MFA*	**MFA****	**univariate*****
c=1	22.92%	0.02%	
c=1,2	24.22%	0.45%	0.04%
c=1,2,3	45.32%	1.49%	
NSNPs=1000, NcSNPs=1000			
	ICA-MFA*	**MFA****	**univariate*****
c=1	22.10%	0.001%	
c=1,2	30.20%	0.004%	0.04%
c=1,2,3	34.97%	0.80%	

Table 3. Percentage of variability explained depending on the number of components, c, included in the model and the number of SNPs. Scenarios from 4 to 6. Legend: N, Number of individuals; NSNPs, Number of SNPs; NcSNPs, Number of causal SNPs; NROIs, Number of brain structures. ICA-MFA. Independent Multifactorial Association method; MFA, Multifactorial Analysis.

N=1000 *NSNPs=100, NcSNPs=10*			
	ICA-MFA*	**MFA****	**univariate*****
c=1	7.19%	0.13%	
c=1,2	8.60%	0.18%	0.04%
c=1,2,3	8.86%	0.35%	
NSNPs=1000, NcSNPs=10			
	ICA-MFA*	**MFA****	**univariate*****
c=1	0.00%	0.19%	
c=1,2	30.55%	0.02%	0.04%
c=1,2,3	30.95%	0.50%	

Table 4. Percentage of variability explained depending on the number of components, c, included in the model and the number of SNPs. Scenarios from 7 to 8. Legend: N, Number of individuals; NSNPs, Number of SNPs; NcSNPs, Number of causal SNPs; NROIs, Number of brain structures. ICA-MFA, Independent Multifactorial Association method; MFA, Multifactorial Analysis.

N=3000			
NSNPs=100, NcSNPs=10			
	ICA-MFA*	MFA**	univariate***
c=1	21.12%	0.00%	
c=1,2	21.44%	0.07%	2.E-05%
c=1,2,3	23.32%	0.14%	
NSNPs=1000, NcSNPs=10			
	ICA-MFA*	MFA**	univariate***
c=1	14.35%	0.29%	
c=1,2	19.10%	2.85%	0.01%
c=1,2,3	22.01%	2.97%	

Table 5. Percentage of variability explained depending on the number of components, c, included in the model and the number of SNPs. Scenarios from 10 to 11. Legend: N, Number of individuals; NSNPs, Number of SNPs; NcSNPs, Number of causal SNPs; NROIs, Number of brain structures. ICA-MFA, Independent Multifactorial Association method; MFA, Multifactorial Analysis.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	PctExp
PC1	1	994	994	20	7.31E-06	0.495%
PC2	1	14	14	0.27	5.99E-01	0.0068%
PC3	1	15	15	0.3	5.80E-01	0.0075%
Residuals	4053	199911	49			99.49%
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	PctExp
IC1	1	18194	18194	448	2.22E-94	9.05%
IC2	1	1165	1165	29	9.03E-08	0.58%
IC3	1	16937	16937	417	2.71E-88	8.43%
Residuals	4053	164638	41			81.94%

Table 6. Variability explained by each principal component (PC1,PC2, PC3; MFA) and by each independent component (IC1, IC2, IC3; ICA-MFA) of the global variance of executive cognitive function. Legend: Df, Degrees of Freedom; Sum Sq, Sum of Squares; Mean Sq, Mean Squares; F value, F ratio; Pr(>F), Pvalue; PctExp, Percentage of variability explained.

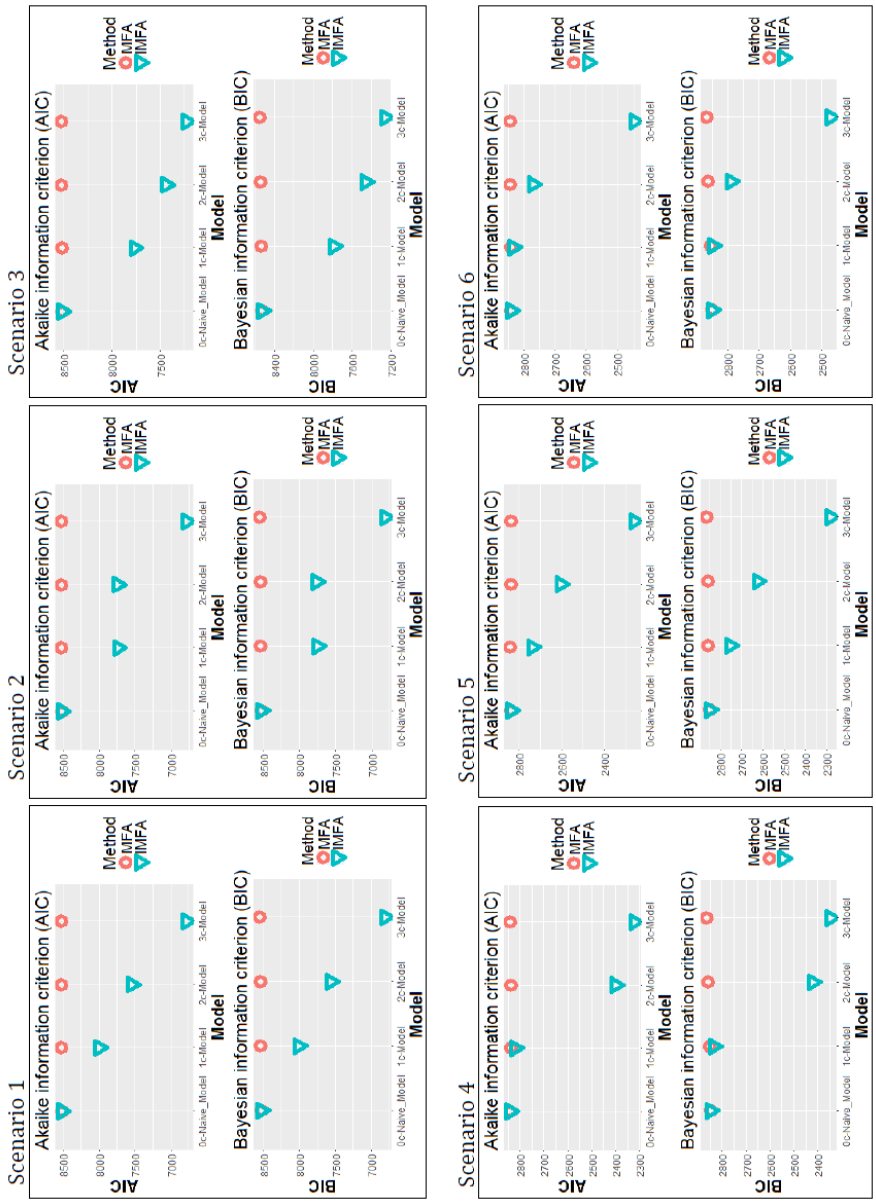

Figure 1. Comparison of the Goodness of fit based on the Akaike information criterion (AIC), considering the inclusion of c=0, 1,2 and 3 components for ICA-MFA and MFA methods.Scenario from 1 to 6. Models with lower AIC values are preferred.

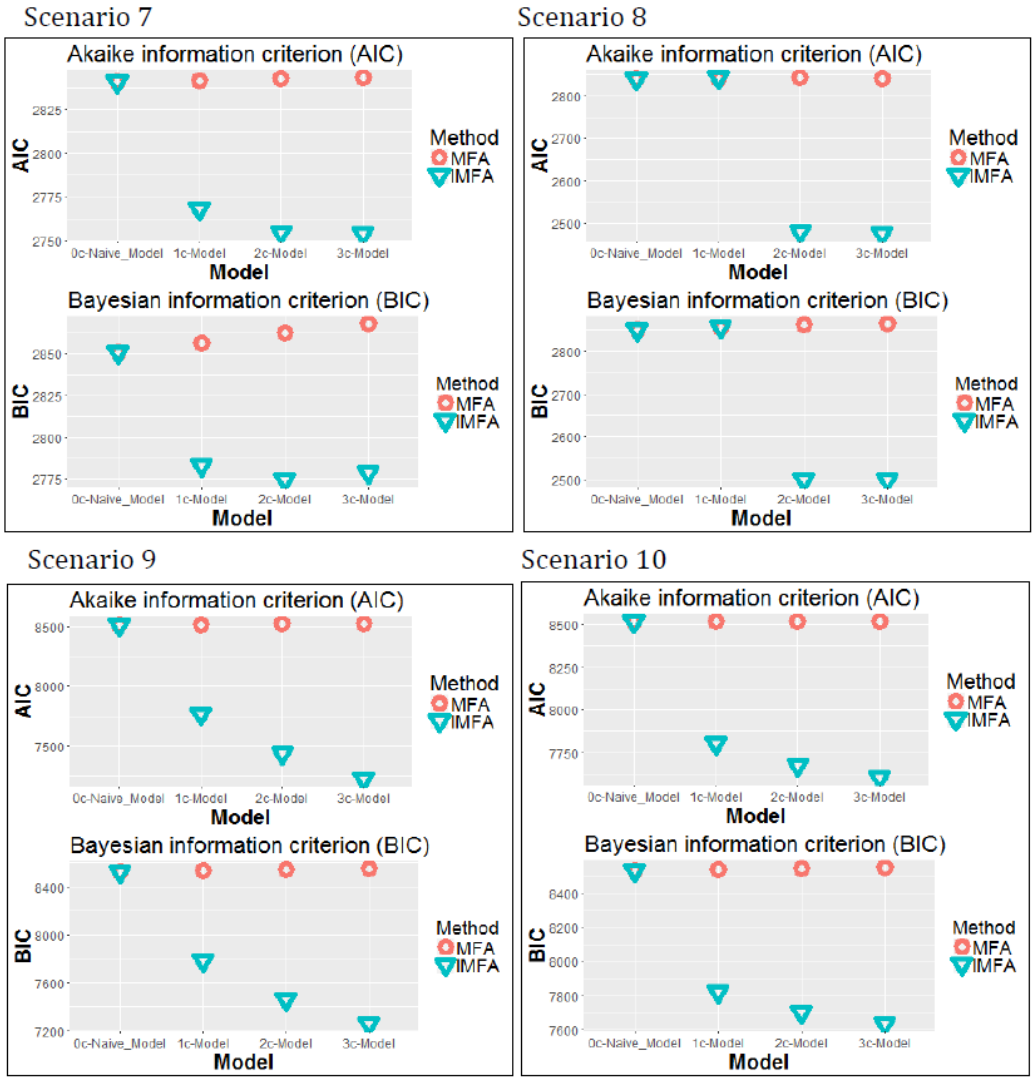

Figure 2. Comparison of the Goodness of fit based on the Akaike information criterion (AIC), considering the inclusion of c=0, 1,2 and 3 components for ICA-MFA and MFA methods. Scenario from 7 to 10.

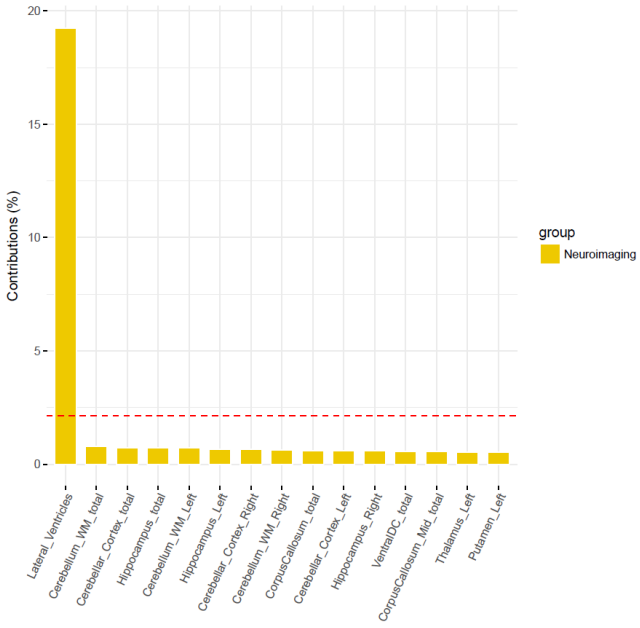

Figure 3. Quantitative features of the study ordered by degree of contribution with the first dimension of Multifactorial Analysis (ICA-MFA).

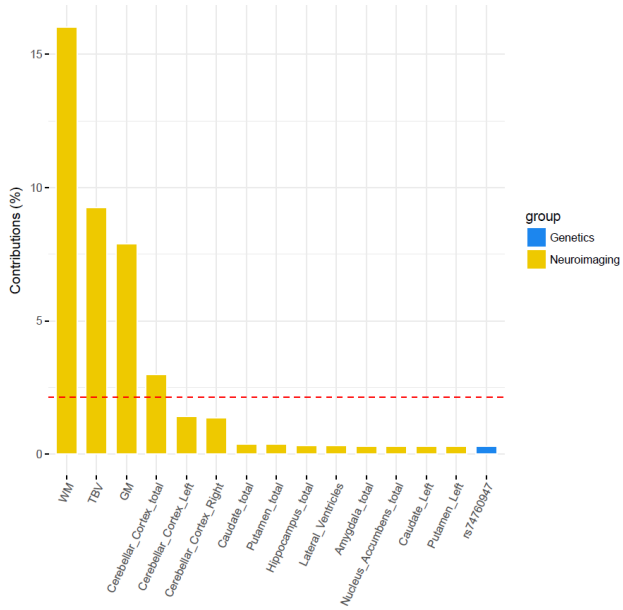

Figure 4. Quantitative features of the study ordered by degree of contribution with the second dimension of Multifactorial Analysis (ICA-MFA).

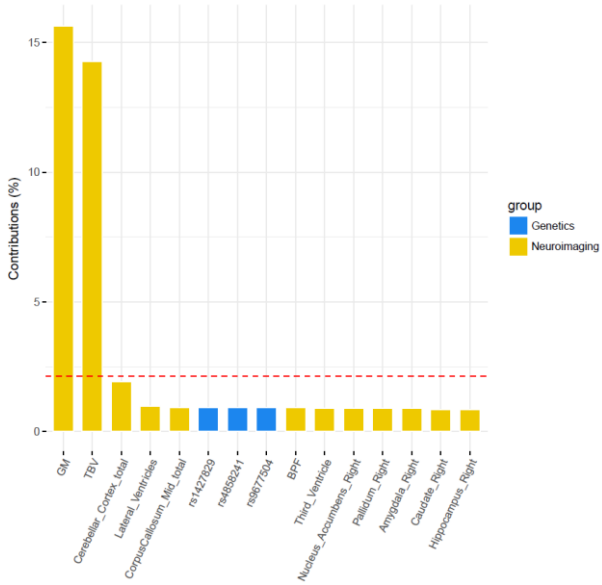

Figure 5. Quantitative features of the study ordered by degree of contribution with the third dimension of Multifactorial Analysis (ICA-MFA).

Supplementary Material

Independent Multifactorial Association Analysis to analyze multiblock data in Imaging Genetics.

Authors: Vilor-Tejedor N, Ikram MA, Roshchupkin GV, Cáceres A, Alemany S, Vernooij MW, Niessen WJ, van Duijn CM, Bustamante M, Pujol J, Sunyer J, Adams HH, González JR.

Table S1. Characteristics of the MRI subsample. Means, SD, Median, and ranges values are shown for continuous variables. *Legend: ICV = Total Intracranial volume; TBV = Total Brain volume; BPF = Brain Parenchymal. Segmentation GM, WM performed with Free Surfer 5.3 image analysis suite.*

Table S2. Characteristics of SNPs associated with ADHD identified in the GWAS meta-analysis from Demontis et al., 2017. *Legend: SNP = Single Nucleotide Polymorphisms; BP = base position; A1 = major allele; A2 = minor allele; MAF = Minor allele frequency.*

Fig S1. Flow chart depicting the final sample size of the real application. Solid lines and boxes represent individuals remaining in the study. Dashed lines and boxes represent individuals excluded. Reason and number of individuals excluded is indicated in dashed boxes.

Fig S2. Quantitative features of the study ordered by degree of contribution with the first dimension of Multifactorial Analysis (MFA).

Fig S3. Quantitative features of the study ordered by degree of contribution with the second dimension of Multifactorial Analysis (MFA).

Fig S4. Quantitative features of the study ordered by degree of contribution with the third dimension of Multifactorial Analysis (MFA).

Sample		N=4057			
Age in years, mean (SD), range		64.7(10.7), 45.0-97.0			
Sex distribution (F/M)(%)		2236/1821 (55.1%)			
Executive Function		30.5(6.6), 3.0-54.0			
Variable		**Mean(SD)**	**Min**	**Median**	**Max**
ICV, mm3		1480310(161558)	926412	1476554	2075800
GM, mm3		597495(58118)	396242	595222	802581
WM, mm3		436338(64199)	216510	435200	684253
TBV, mm3		1033834(117791)	640449	1030498	1440285
BPF, %		0.7 (0.05)	0.51	0.7	0.92
CSF, mm3		1259(354)	504	1192	3895
Ventricles, mm3					
	Lateral	26016(16100)	2900	21516	146082
	Third	1533(641)	459	1389	4822
	Fourth	1994(569)	454	1906	7865
Hippocampus, mm3					
	Total	7739(1060)	2191	7826	11537
	Right	3877(577)	411	3928	5734
	Left	3862(615)	627	3928	6687
Nucleus Accumbens, mm3					
	Total	1054(173)	264	1045	1829
	Right	492(94)	87	486	906
	Left	561(97)	177	555	1168
Amygdala, mm3					
	Total	2718(389)	1045	2713	4448
	Right	1402(216)	113	1399	2394
	Left	1317(212)	304	1315	2255
Caudate Nucleus, mm3					
	Total	6889(1080)	3454	6742	14248
	Right	3502(562)	1620	3437	7613
	Left	3387(541)	1834	3315	7155
Globus Pallidus, mm3					
	Total	2904(468)	1508	2879	6318
	Right	1423(255)	481	1406	3382
	Left	1481(238)	499	1475	2936
Putamen, mm3					
	Total	9112(1257)	3965	9019	17397
	Right	4474(643)	1663	4438	8426
	Left	4638(653)	1929	4592	9035
Thalamus, mm3					
	Total	12578(1566)	7178	12454	26764
	Right	6294(796)	3310	6239	11937
	Left	6285(794)	3592	6231	14827
Cerebellum WM, mm3					
	Total	23737(3497)	6398	23505	44160
	Right	11870(1794)	2947	11751	21065
	Left	11868(1802)	3451	11753	26288
Cerebellum Cortex, mm3					
	Total	99295(10930)	34039	99099	143580
	Right	50251(5680)	17330	50126	72111
	Left	49044(5459)	16709	48943	72696
Corpus callosum, mm3		2639(539)	478	2664	4601
Corpus callosum Mid, mm3		1056(266)	131	1045	2199
Ventral DC, mm3		7701(877)	4904	7650	11744

Table S1. Characteristics of the MRI subsample. Means, SD, Median, and ranges values are shown for continuous variables. Legend: ICV = Total Intracranial volume; TBV = Total Brain volume; BPF = Brain Parenchymal. Segmentation GM, WM performed with Free Surfer 5.3 image analysis suite.

SNP	CHR	BP	A1	A2	MAF	Gene
rs9677504	2	215181889	G	A	0.093	*SPAG16*
rs4858241	3	20669071	T	G	0.386	*Intergenic*
rs4916723	5	87854395	A	C	0.436	*LINC00461, MIR9-2, LINC02060*
rs74760947	8	34352610	A	G	0.041	*LINC01288*
rs11591402	10	106747354	T	A	0.221	*SORCS3*
rs1427829	12	89760744	A	G	0.436	*DUSP6, POC1B*
rs281324	15	47754018	C	T	0.474	*SEMA6D*
rs212178	16	72578131	A	G	0.103	*LINC01572*

Table S2. Characteristics of SNPs associated with ADHD identified in the GWAS meta-analysis from Demontis et al., 2017. Legend: SNP = Single Nucleotide Polymorphisms; BP = base position; A1 = major allele; A2 = minor allele; MAF = Minor allele frequency.

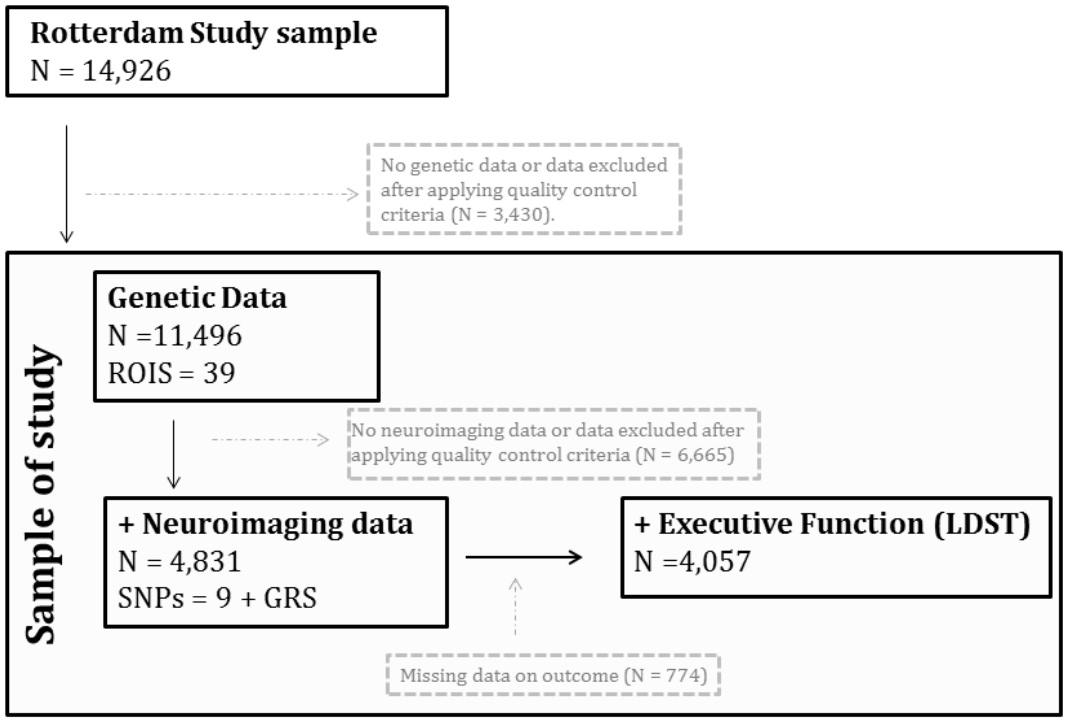

Figure S1. Flow chart depicting the final sample size for each outcome analyzed. Solid lines and boxes represent individuals remaining in the study. Dashed lines and boxes represent individuals excluded. Reason and number of individuals excluded is indicated in dashed boxes.

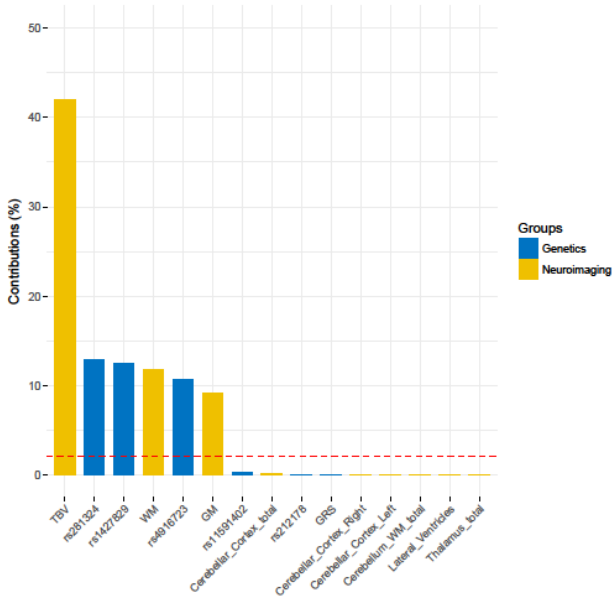

Figure S2. Quantitative features of the study ordered by degree of contribution with the first dimension of Multifactorial Analysis (MFA).

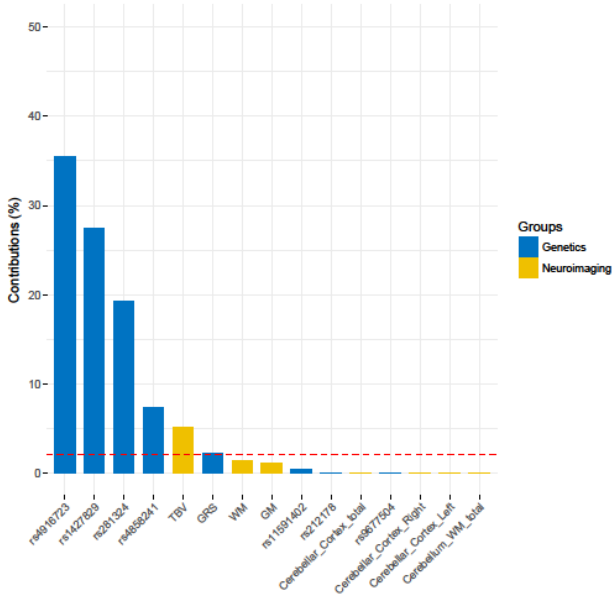

Figure S3. Quantitative features of the study ordered by degree of contribution with the second dimension of Multifactorial Analysis (MFA).

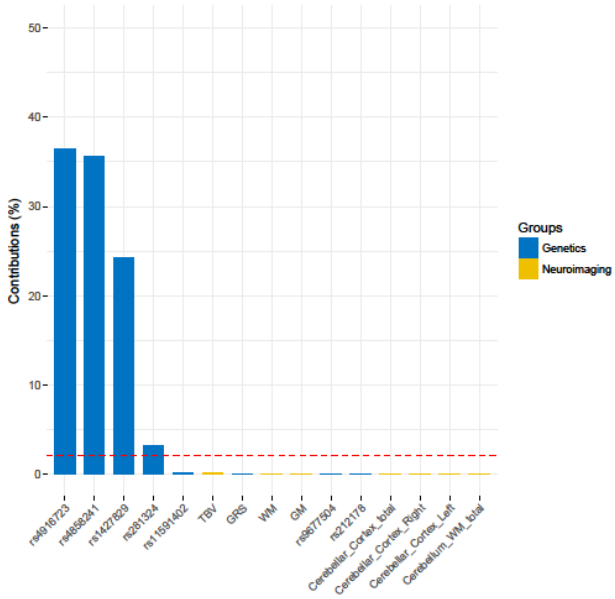

Figure S4. Quantitative features of the study ordered by degree of contribution with the third dimension of Multifactorial Analysis (MFA).

Chapter 6

Longitudinal assessment of genetic risk factors on brain structure

Genome-wide genetic risk variants for ADHD predict longitudinal changes of ventrical structures in a population-based longitudinal sample

Authors: **Vilor-Tejedor N**, Ikram MA, Roshchupkin G, Nieessen WG, Alemany S, Adams HH.

(in preparation)

Title: Genome-wide genetic risk variants for ADHD predicts longitudinal changes of ventricle structures in a population-based longitudinal sample

Authors: **Vilor-Tejedor N*** (1-3); Ikram MA (4-6); Roshchupkin G (4,7); Niessen WG (5,7-8); Alemany S (1-3); Adams HH (4,5).

(1) ISGlobal - Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

(2) Universitat Pompeu Fabra (UPF), Barcelona, Spain.

(3) CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

(4) Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands

(5) Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands

(6) Department of Neurology, Erasmus MC, Rotterdam, the Netherlands

(7) Department of Medical Informatics, Erasmus MC, Rotterdam, the Netherlands

(8) Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands

*Correspondence to: Natàlia Vilor-Tejedor, Barcelona Institute for Global Health (ISGlobal). C. Doctor Aiguader 88, 08003 Barcelona, Spain. E-mail: natalia.vilor@isglobal.org

ORCID: 0000-0003- 4935-6721

CONFLICT OF INTEREST

The authors declare no conflicts of interest and financial disclosures.

Abstract

Background. Brain development during childhood and adolescence generally involves a decrease of cortical grey matter volume (GM) and an increase of cerebral white matter volume (WM). However, there are still inconsistencies regarding the developmental pattern of brain changes during adulthood and few studies assessing these issues. Longitudinal designs of brain measures acquired through magnetic resonance imaging (MRI) are emerging for the understanding of typical development as well as neurodevelopmental domains. Since brain structural abnormalities have been linked to Attention-Deficit/Hyperactivity disorder (ADHD), and brain structural development is genetically influenced, we may expect that genes related to ADHD may have a role in brain structural development

Objective. To investigate associations between genome-wide significant SNPs reported for ADHD and degenerative trajectories of brain structures during adulthood. To compare the results with those from cross-sectional designs.

Design. Longitudinal and cross-sectional imaging genetic designs in a representative population-based sample of adults (Rotterdam study). 3220 individuals, aged from 45 to 100 years old.

Main outcomes. Brain volumes derived from structural MRI scans using automated tissue segmentation.

Results. We observed significant volumetric differences in nonlinear trajectories in the ventricle structures (lateral ventricles and third ventricle) related to rs212178 genotypic variation. In addition, we observed significant differences in nonlinear trajectories in the amygdala structure related to rs4916723, and significant differences in nonlinear trajectories in caudate structure related to rs281324.

Conclusion and Relevance. Genetic risk variants for ADHD predict developmental decreases at the ventricles, amygdala and caudate regions during late adulthood. These trajectories involve non-linear dynamics in parts of the brain that have been extensively discussed in the context of cognition, psychiatric and neurodegenerative diseases.

Keywords: Brain structure trajectories; Imaging genetics; neuroimaging, ventricle structures, genetics.

1. INTRODUCTION

Attention-deficit hyperactivity disorder (ADHD) is a prevalent childhood neurodevelopmental disorder with an estimated worldwide prevalence of 5.2% (Willcutt 2012; Thomas et al. 2015; Faraone et al. 2015). Although it is considered most common in children, recent work suggests that for some individuals, ADHD first emerges in adulthood. In addition, one sixth of individuals with a childhood diagnosis continue to meet clinical criteria for ADHD in adulthood (Faraone, Biederman, and Mick 2006; Moffitt et al. 2015; Agnew-Blais et al. 2016; Caye, Rocha, et al. 2016). The determinants of the persistence of ADHD are not fully understood, although cortical maturation, deficiencies on dopaminergic and serotonergic pathways (Bralten et al. 2013), comorbidities (Kessler et al. 2005; Torres et al. 2017; Noordermeer, Luman, and Oosterlaan 2016; Leaberry et al. 2017) and other factors have been considered as contributors (Fayyad et al. 2017; Caye, Spadini, et al. 2016).

The most promising genetic findings were reported in a recent GWAS of ADHD performed by the ADHD working group of the Psychiatric Genomics Consortium (PGC) and the Early Genetics and Lifecourse Epidemiology (EAGLE) Consortium (Demontis et al. 2017). The study including 20,183 ADHD cases and 35,191 controls identified, for the first time, 12 genome-wide significant independent loci (Pvalue $< 10^{-8}$). However, the directly understanding of genetic contributions is starting to report reliable results and most of the studies are starting to focus on composite measures, such as genetic scores, to assess the polygenic architecture of ADHD (Thapar et al. 1999; Hamshere et al. 2013; Martin et al. 2015). Furthermore, given the highly heritable nature of ADHD, studies of intermediate phenotypes such as brain measures that determine genetic liability seem particularly called for targeting genetic effects for ADHD (Rommelse et al. 2011).

Neuroanatomical findings reported for ADHD include reductions at grey matter (GM) whole-brain, cortical and subcortical (Greven et al. 2015; Hoogman et al. 2017; Frodl and Skokauskas 2012; Norman et al. 2016; Noordermeer et al. 2017). Specifically, GM reductions have been reported in specific regions such as, reduced cortical thickness of the frontal and temporal lobes and occipital areas (Schweren et al. 2015; Kumar, Arya, and Agarwal 2017; Ambrosino et al. 2017). A potential explanation for the inconsistent neuroanatomical findings for ADHD could be the cross-sectional assessment of most of the studies. So far, there have not been conducted studies on neuroanatomical correlates exclusively focused on longitudinal changes influenced by genetic components. This is an

important gap in the performance of designs due to cross-sectional designs can only capture the mean differences of a phenotype across the subgroups of subjects. In contrast, longitudinal models allow us to estimate not only the mean values of a quantitative trait (QT) at the baseline, but also the rates of QT changes into each group depending on age/ including a group-age interaction. Thus, compared to the cross-sectional design, the longitudinal design can provide increased statistical power by reducing the confounding effect of between-subject variability and provide unique insights into the temporal dynamics of the underlying biological process of neurodevelopmental domains (Bernal-Rusiel et al. 2013). Moreover, current research suggests that the effects of shared genetic liability may have longitudinal effects on brain structure (Lee et al. 2016; Gu and Kanai 2014). However, there is still few studies addressing the combined relationship between longitudinal changes on brain structure and genetics on ADHD.

Since brain structural abnormalities have been linked to ADHD, and brain structural development is genetically influenced, we may expect that genes related to ADHD may have a role in brain structural development. Hence, the current study aimed to disentangle longitudinal brain regional volume changes associated with risk genetic variants for ADHD. We examined the associations between genome-wide significant SNPs reported for ADHD and population-based trajectories of brain structures during adulthood. Moreover, due to the polygenic architecture of ADHD, where common risk alleles have small effect sizes, we also inspected associations between a composite genetic risk score (GRS) with brain structure trajectories.

2. METHODS AND MATERIALS

2.1 Study Population

Data used in the preparation of this article were drawn from the Rotterdam Study. The Rotterdam Study is an ongoing population-based cohort study in Netherlands, which currently consists of 14,926 individuals aged 45 years or more at baseline (Ikram et al. 2017). A total of 5430 individuals were scanned through magnetic resonance imaging (MRI) and eligible for this study. Individuals with incomplete MRI-acquisitions, scans with artifacts and dementia or stroke were excluded. This resulted in a final study population of 4,831 non-demented individuals with information also available on genotyping data [Figure 1]. The

Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study). All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

2.2 Image acquisition, processing and selection.

Magnetic Resonance Imaging scanning was done on a 1.5-T MRI scanner (Signa Excite II; General Electric Healthcare, Milwaukee, WI, USA). The MRI protocol included a high-resolution axial T1-weighted 3-dimensional fast radio frequency spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse (FASTSPGR-IR) sequence (repetition time [TR] = 13.8 ms, echo time [TE] = 2.8 ms, inversion time [TI] = 400 ms, field of view [FOV] = 25 cm2, matrix = 416 × 256, flip angle = 20°, number of excitations [NEX] = 1, bandwidth [BW] = 12.50 kHz, 96 slices with slice thickness 1.6 mm 0-padded to 0.8 mm). All slices were contiguous. According to the Rotterdam Study standard acquisition protocol images were resampled to 512 x 152 x 192 voxels (voxel size: 0.5 x 0.5 x 0.8 mm3). The T1-weighted MRI scans were processed using a model-based automated procedure of Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) (Fischl et al. 2004) to obtain segmentations and volumetric summaries of subcortical structures and thickness of the cerebral cortex. This procedure automatically assigns a neuroanatomical label to each voxel in an MRI volume based on probabilistic information obtained from a manually labeled training set. After quality control, a total of 12,174 brain MR-scans have been obtained in over 5800 individuals, as of July 2015. From the total sample of individuals, 4831 individuals have one scan measurement, 3220 have at least two scan measurements, 1887 have at least three scan measurements, and 141 have four scan measurements. Further details of the MRI protocol can be found in Ikram et al. 2015.

2.3 Genotyping acquisition and genetic variant selection

The Illumina 550K, 550K duo and 610K arrays were used for genotyping Samples with a call rate below 97.5%, gender mismatch, excess autosomal heterozygosity (>0.336), duplicates or family relations and ethnic outliers were excluded. Genetic variants were filtered by Hardy-Weinberg equilibrium ($P<10^{-6}$), allele frequency (excluding minor allele frequency (MAF < 0.001) and SNP call rate with a minimum of 98%. Genotypes were imputed using

MACH/minimac software to the 1000 Genomes phase I version 3 reference panel (all populations). Among the variants, a total of 9 SNPs associated with ADHD at a genome-wide threshold of significance ($P<10^{-8}$), were pre-selected [Table SM1] (Demontis et al. 2017). Furthermore, we constructed a genetic risk score (GRS) by multiplying the number of risk alleles by their reported odds ratio (after natural logarithm transformation) for the disease and summing this weighted allele score of each variant up into a disease risk score for ADHD.

2.5 Statistical Analysis

We use a linear mixed-effects (lme) model with a random slope and a random intercept to calculate trajectories of individual subject. We tested the longitudinal association between genome-wide genetic risk variant for ADHD and brain structures in fully adjusted models . A total of 3,320 individuals with at least two repeated measures of MRI-scan were included, resulting in 8468 observations in total. Moreover, to robustly describe the non-linearity of the effects of age, we included a polynomial interpolation, considering two predefined spline terms:

$$Y_{ik} = (\gamma_{00} + U_{0i}) + (\gamma_{10} + U_{1i})t_{ik} + \sum_{j=1}^{g-1} X_{ij}\beta_j + SNP_i\beta_g + \big(SNP_i \times ns(t_{ik}, 2)\big)\beta_{j+g+1} + \varepsilon_{ik},$$

where Y_{ik} is the regional imaging phenotype of subject $i = 1, 2, \ldots, N$ at the k-th measurement; $t_{ik} = k$ represents age at the time point k; and U_{0i} and $U_{1i}$are covariate vectors of i-th subject at k-th measurement for random effects γ_{00}, $\gamma_{10} \sim N(0, D)$. In addition, X_{ij} represents the covariate matrix of i-th subject at k-th measurement for fixed effects β_j that include: sex, and ICV; $SNP_i = 0$, 1 or 2 is the count of the risk allele for the SNPs to be tested; and $e_{ik} \sim N(0, \sigma^2)$ is an independent error term. Age-by-genotype interactions for each regional volume were included in the lme. The coefficients of the interaction terms quantified the longitudinal change in brain volume and how genotypes of each SNP modified the magnitude of these changes.

Further, general linear models (GLM) tested cross-sectional differences in regional volumes between genotypes as a function of age using only the data corresponding to the first scan-acquisition of the sample (first observation per subject), as would be the case in a simple single observation cross-sectional study.

The standard GLM was defined as:

$$Y_i = \beta_0 + \sum_{j=1}^{3} X_{ij}\beta_j + SNP_i\beta_4 + \varepsilon_i,$$

where Y_i is the regional imaging phenotype of subject $i = 1, 2, \ldots, N$ at the baseline. $SNP_i = 0$, 1 or 2 is the count of the minor allele for the SNPs to be tested; X_{ij} represents the covariate matrix which include: sex, age and baseline intracranial volume (ICV); and $\varepsilon_i \sim N(0, \sigma^2)$ is an independent error term.

Weighted effects of GLM and lme were corrected for multiple comparisons by false discovery rate (FDR 5%) method. The effective number of independent tests, Me, was estimated under a permutation procedure, assuming B=10,000 permutations. The resulting adjusted threshold of significance was set on 3E-03. All statistical analyses were carried out using R.

3. RESULTS

3.1 Descriptive results

Basic descriptive characteristics of the subjects with at least two valid MRIs, mean and range of age, and sex ratios for each MRI-scan acquisition and descriptive characteristics of the brain volumetric data for each scan acquisition, are presented in Table 1. The study included in total, 2659 women and 2172 men between 45 and 100 years of age. The distribution in the percentage of women/men is balanced in all the scan acquisitions, while the age becomes slightly higher in the last visit, as expected.

3.2 Cross-section association results

Main genetic effects

Table 2 showed the adjusted SNP coefficients for brain structure volumes. No significant results after multiple comparison correction were found. At nominal level of significance, we found that the risk allele of the SNP rs212178 was negatively associated with cerebellar cortex, Gray Matter (GM), BPF and Thalamus. Also, we found a negative effect of rs4916723, and a positive effect of rs4858241 in the fourth ventricle. In addition, we found a

positive association in corpus callosum for the risk allele of the SNP-rs11591402, and the risk allele of the SNP-rs281324.

*Age*genetic interaction effects*

Figure 2 showed main results for the adjusted Age*SNP coefficients considering two spline effects for brain structure volumes. For the first spline term, we found significant negative interaction effects with ventricle structures (lateral ventricles, P=0.01; and fourth ventricle, P=0.009) and a positive effect with GM (P=0.005) related to rs212178 genotypic variations. We also observed a negative significant interaction effect with Cerebellum White Matter (P=0.04) related to rs4858241 genotypic variations, a negative significant interaction effect with Thalamus (P=0.04) and Ventral DC (P=0.012) related to rs4916723 genotypic variations, and also a negative significant interaction effect with White Matter (P=0.023) and Corpus Callosum (P=0.018), and a positive significant effect with Lateral ventricles (P=0.01) related to rs74760947 genotypic variations.

For the second spline term, we found significant negative interaction effects with ventricle structures (lateral ventricles, P= 0.03 ; Fourth Ventricle, P= 0.041: Third Ventricle, P= 0.008), and also with Caudate (P=0.04) related to rs212178 genotypic variations. Moreover, we observed a negative significant interaction effect with GM (P=0.03), BPF (P=0.02), CSF (P=0.03), Third ventricle (P=0.03) and Cerebellum WM (P=0.02), related to rs4858241 genotypic variations. In addition, we found a negative significant interaction effect with Putamen (P=0.01) related to rs1427829, a negative significant interaction effect with Hippocampus (P= 0.04) and Corpus Callosum (P= 0.04) related to rs281324, and a positive significant interaction effect with Globus Pallidus (P=0.02), related to rs9677504. Moreover, the results suggest the existence of differences in the change in effect at each spline term [Figure 2]. Results do not remain significant after multiple comparison correction (FDR correction with $\alpha = .05$, required adjusted p-value $\leq$ 3E-03). Furthermore, no Age*GRS effects were found on either brain structures of interest.

3.3 Longitudinal association results

Mixed-models analysis revealed significant age-by-SNP interactions. Specifically, we observed significant differences in nonlinear trajectories in the ventricle structures (Lateral Ventricles, p= 4.1E-05; Third Ventricle, p= 2.5E-03; Fourth ventricle, p=5.5E-03) related to rs212178 genotypic variations [Figures 3-4]. In addition, we observed significant differences

in nonlinear trajectories in the amygdala structure related to rs4916723 (p=0.0014), and significant differences in nonlinear trajectories in caudate structure related to rs281324 (p=0.0018).

4. DISCUSSION

Brain structural development is genetically influenced (refs Silvi), but genetic basis remains still largely unknown. However, few studies in IG have assessed genetic effects on longitudinal brain changes, even though structural changes that occur during development and ageing are related to mental health and general cognitive functioning. On the other hand, individuals affected by ADHD present widespread structural abnormalities. Furthermore, ADHD is a highly heritable neurodevelopment disorder, onset in childhood (Thapar et al. 1999), and GWAS hits have been recently identified (Demontis et al. 2017).

Taking all together, we hypothesized that the longitudinal effects on brain structure related to ADHD are likely to be assessed through effects of ADHD genetic risk factors.

According to our knowledge this is the first study aimed to investigate genetic effects of genome-wide significant SNPs for ADHD on longitudinal brain changes. Our main finding was that rs212178 genotypes were associated with less decline in ventricle brain volumes over time. Changes on brain ventricle structures have been significantly associated with poor performance on computerized tests that assess executive function and cortical function (Best and Miller 2010). However, factors that contribute to the process of cerebral atrophy underlying enlargement of the ventricle structures remain to be answered. Vascular factors seem to be implicated in a substantial proportion of patients with dilatation of the ventricle system However, this does not explain the relation between ventricular enlargement and cognitive function. In this study, we focus on the analysis of genetic factors underlying the involvement of non-linear dynamics in parts of the brain that have been extensively discussed in the context of cognition (Hoth et al. 2010; Carmichael et al. 2007), Alzheimer's disease (Nestor et al. 2008; Thompson et al. 2004), and Parkinson's disease(Mak et al. 2017). Moreover, we suggest insights into the genetic determinants of structural brain changes that may contribute to cognitive impairments in later life.

The main strength of this study was the ability to examine intraindividual change in brain structures using longitudinal data, which include at least 2 scanner acquisitions per individual. This provides a more valid measure of the brain structural change than extrapolating an estimate of change from separate individuals across a range of ages using cross-sectional designs. Moreover, we considered the statistical interaction between genetic variations and age because of the implications for the shape of the distribution of onset age in risk analyses.

Further research may greatly benefit from longitudinal designs which represent a potential form to increase the statistical power to detect significant causal factors affecting structural brain changes.

ACKNOWLEDGEMENTS

Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2017 FI_B 00636), Generalitat de Catalunya – Fons Social Europeu. This work has been partially supported by a STSM Grant from EU COST Action 15120 Open Multiscale Systems Medicine (OpenMultiMed) and Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP). ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

Silvia Alemany thanks the Institute of Health Carlos III for her Sara Borrell postdoctoral grant (CD14/00214).

The generation and management of GWAS genotype data for the Rotterdam Study are supported by the Netherlands Organization of Scientific Research NWO Investments (no. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project no. 050-060-810. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. This research is supported by the Dutch Technology Foundation STW (12723), which is part of the NWO, and which is partly funded by the Ministry of Economic Affairs.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project: ORACLE, grant agreement No: 678543).

REFERENCES

Agnew-Blais, Jessica C., Guilherme V. Polanczyk, Andrea Danese, Jasmin Wertz, Terrie E. Moffitt, and Louise Arseneault. 2016. "Evaluation of the Persistence, Remission, and Emergence of Attention-Deficit/Hyperactivity Disorder in Young Adulthood." *JAMA Psychiatry* 73 (7): 713. doi:10.1001/jamapsychiatry.2016.0465.

Ambrosino, Sara, Patrick de Zeeuw, Lara Marise Wierenga, Sarai van Dijk, and Sarah Durston. 2017. "What Can Cortical Development in Attention-Deficit/Hyperactivity Disorder Teach Us About the Early Developmental Mechanisms Involved?" *Cerebral Cortex* 27 (9): 4624–34. doi:10.1093/cercor/bhx182.

Bernal-Rusiel, Jorge L, Douglas N Greve, Martin Reuter, Bruce Fischl, Mert R Sabuncu, and Alzheimer's Disease Neuroimaging Initiative. 2013. "Statistical Analysis of Longitudinal Neuroimage Data with Linear Mixed Effects Models." *NeuroImage* 66 (February). NIH Public Access: 249–60. doi:10.1016/j.neuroimage.2012.10.065.

Best, John R, and Patricia H Miller. 2010. "A Developmental Perspective on Executive Function." *Child Development* 81 (6). NIH Public Access: 1641–60. doi:10.1111/j.1467-8624.2010.01499.x.

Bralten, Janita, Barbara Franke, Irwin Waldman, Nanda Rommelse, Catharina Hartman, Philip Asherson, Tobias Banaschewski, et al. 2013. "Candidate Genetic Pathways for Attention-Deficit/hyperactivity Disorder (ADHD) Show Association to Hyperactive/impulsive Symptoms in Children with ADHD." *Journal of the American Academy of Child and Adolescent Psychiatry* 52 (11): 1204–1212.e1. doi:10.1016/j.jaac.2013.08.020.

Carmichael, Owen T., Lewis H. Kuller, Oscar L. Lopez, Paul M. Thompson, Rebecca A. Dutton, Allen Lu, Sharon E. Lee, et al. 2007. "Cerebral Ventricular Changes Associated With Transitions Between Normal Cognitive Function, Mild Cognitive Impairment, and Dementia." *Alzheimer Disease & Associated Disorders* 21 (1): 14–24. doi:10.1097/WAD.0b013e318032d2b1.

Caye, Arthur, Thiago Botter-Maio Rocha, Luciana Anselmi, Joseph Murray, Ana M. B. Menezes, Fernando C. Barros, Helen Gonçalves, et al. 2016. "Attention-Deficit/Hyperactivity Disorder Trajectories From Childhood to Young Adulthood." *JAMA Psychiatry* 73 (7): 705. doi:10.1001/jamapsychiatry.2016.0383.

Caye, Arthur, Alex V. Spadini, Rafael G. Karam, Eugenio H. Grevet, Diego L Rovaris, Claiton H. D. Bau, Luis A. Rohde, and Christian Kieling. 2016. "Predictors of

Persistence of ADHD into Adulthood: A Systematic Review of the Literature and Meta-Analysis." *European Child & Adolescent Psychiatry* 25 (11): 1151–59. doi:10.1007/s00787-016-0831-8.

Demontis, Ditte, Raymond K. Walters, Joanna Martin, Manuel Mattheisen, Thomas Damm Als, Esben Agerbo, Rich Belliveau, et al. 2017. "Discovery Of The First Genome-Wide Significant Risk Loci For ADHD." *bioRxiv*, June. Cold Spring Harbor Laboratory, 145581. doi:10.1101/145581.

Faraone, Stephen V., Philip Asherson, Tobias Banaschewski, Joseph Biederman, Jan K. Buitelaar, Josep Antoni Ramos-Quiroga, Luis Augusto Rohde, Edmund J. S. Sonuga-Barke, Rosemary Tannock, and Barbara Franke. 2015. "Attention-Deficit/hyperactivity Disorder." *Nature Reviews Disease Primers* 1 (August). Nature Publishing Group: 15020. doi:10.1038/nrdp.2015.20.

Faraone, Stephen V, Joseph Biederman, and Eric Mick. 2006. "The Age-Dependent Decline of Attention Deficit Hyperactivity Disorder: A Meta-Analysis of Follow-up Studies." *Psychological Medicine* 36 (2): 159–65. doi:10.1017/S003329170500471X.

Fayyad, John, Nancy A. Sampson, Irving Hwang, Tomasz Adamowski, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Laura H. S. G. Andrade, et al. 2017. "The Descriptive Epidemiology of DSM-IV Adult ADHD in the World Health Organization World Mental Health Surveys." *ADHD Attention Deficit and Hyperactivity Disorders* 9 (1): 47–65. doi:10.1007/s12402-016-0208-3.

Fischl, Bruce, David H. Salat, André J.W. van der Kouwe, Nikos Makris, Florent Ségonne, Brian T. Quinn, and Anders M. Dale. 2004. "Sequence-Independent Segmentation of Magnetic Resonance Images." *NeuroImage* 23 (January): S69–84. doi:10.1016/j.neuroimage.2004.07.016.

Frodl, T., and N. Skokauskas. 2012. "Meta-Analysis of Structural MRI Studies in Children and Adults with Attention Deficit Hyperactivity Disorder Indicates Treatment Effects." *Acta Psychiatrica Scandinavica* 125 (2): 114–26. doi:10.1111/j.1600-0447.2011.01786.x.

Greven, Corina U, Janita Bralten, Maarten Mennes, Laurence O'Dwyer, Kimm J E van Hulzen, Nanda Rommelse, Lizanne J S Schweren, et al. 2015. "Developmentally Stable Whole-Brain Volume Reductions and Developmentally Sensitive Caudate and Putamen Volume Alterations in Those with Attention-Deficit/hyperactivity Disorder and Their Unaffected Siblings." *JAMA Psychiatry* 72 (5): 490–99. doi:10.1001/jamapsychiatry.2014.3162.

Gu, Jenny, and Ryota Kanai. 2014. "What Contributes to Individual Differences in Brain Structure?" *Frontiers in Human Neuroscience* 8. Frontiers Media SA: 262. doi:10.3389/fnhum.2014.00262.

Hamshere, Marian L., Kate Langley, Joanna Martin, Sharifah Shameem Agha, Evangelia Stergiakouli, Richard J.L. Anney, Jan Buitelaar, et al. 2013. "High Loading of Polygenic Risk for ADHD in Children With Comorbid Aggression." *American Journal of Psychiatry* 170 (8): 909–16. doi:10.1176/appi.ajp.2013.12081129.

Hoogman, Martine, Janita Bralten, Derrek P Hibar, Maarten Mennes, Marcel P Zwiers, Lizanne S J Schweren, Kimm J E van Hulzen, et al. 2017. "Subcortical Brain Volume

Differences in Participants with Attention Deficit Hyperactivity Disorder in Children and Adults: A Cross-Sectional Mega-Analysis." *The Lancet. Psychiatry* 4 (4). Elsevier: 310–19. doi:10.1016/S2215-0366(17)30049-4.

Hoth, Karin F, Athena Poppas, Kristin E Ellison, Robert H Paul, Andrew Sokobin, Youngsoo Cho, and Ronald A Cohen. 2010. "Link between Change in Cognition and Left Ventricular Function Following Cardiac Resynchronization Therapy." *Journal of Cardiopulmonary Rehabilitation and Prevention* 30 (6). NIH Public Access: 401–8. doi:10.1097/HCR.0b013e3181e1739a.

Ikram, M. Arfan, Guy G. O. Brusselle, Sarwa Darwish Murad, Cornelia M. van Duijn, Oscar H. Franco, André Goedegebure, Caroline C. W. Klaver, et al. 2017. "The Rotterdam Study: 2018 Update on Objectives, Design and Main Results." *European Journal of Epidemiology* 32 (9): 807–50. doi:10.1007/s10654-017-0321-4.

Ikram, M. Arfan, Aad van der Lugt, Wiro J. Niessen, Peter J. Koudstaal, Gabriel P. Krestin, Albert Hofman, Daniel Bos, and Meike W. Vernooij. 2015. "The Rotterdam Scan Study: Design Update 2016 and Main Findings." *European Journal of Epidemiology* 30 (12): 1299–1315. doi:10.1007/s10654-015-0105-7.

Kessler, Ronald C, Lenard A Adler, Russell Barkley, Joseph Biederman, C Keith Conners, Stephen V Faraone, Laurence L Greenhill, et al. 2005. "Patterns and Predictors of Attention-Deficit/hyperactivity Disorder Persistence into Adulthood: Results from the National Comorbidity Survey Replication." *Biological Psychiatry* 57 (11): 1442–51. doi:10.1016/j.biopsych.2005.04.001.

Kumar, Uttam, Amit Arya, and Vivek Agarwal. 2017. "Neural Alterations in ADHD Children as Indicated by Voxel-Based Cortical Thickness and Morphometry Analysis." *Brain and Development* 39 (5): 403–10. doi:10.1016/j.braindev.2016.12.002.

Leaberry, Kirsten D., Paul J. Rosen, Nicholas D. Fogleman, Danielle M. Walerius, and Kelly E. Slaughter. 2017. "Comorbid Internalizing and Externalizing Disorders Predict Lability of Negative Emotions Among Children With ADHD." *Journal of Attention Disorders*, October, 108705471773464. doi:10.1177/1087054717734647.

Lee, P H, J T Baker, A J Holmes, N Jahanshad, T Ge, J-Y Jung, Y Cruz, et al. 2016. "Partitioning Heritability Analysis Reveals a Shared Genetic Basis of Brain Anatomy and Schizophrenia." *Molecular Psychiatry* 21 (12). NIH Public Access: 1680–89. doi:10.1038/mp.2016.164.

Mak, Elijah, Li Su, Guy B Williams, Michael J Firbank, Rachael A Lawson, Alison J Yarnall, Gordon W Duncan, et al. 2017. "Longitudinal Whole-Brain Atrophy and Ventricular Enlargement in Nondemented Parkinson's Disease." *Neurobiology of Aging* 55 (July). Elsevier: 78–90. doi:10.1016/j.neurobiolaging.2017.03.012.

Martin, J, M C O'Donovan, A Thapar, K Langley, and N Williams. 2015. "The Relative Contribution of Common and Rare Genetic Variants to ADHD." *Translational Psychiatry* 5 (2). Nature Publishing Group: e506–e506. doi:10.1038/tp.2015.5.

Moffitt, Terrie E., Renate Houts, Philip Asherson, Daniel W. Belsky, David L. Corcoran, Maggie Hammerle, HonaLee Harrington, et al. 2015. "Is Adult ADHD a Childhood-Onset Neurodevelopmental Disorder? Evidence From a Four-Decade Longitudinal Cohort Study." *American Journal of Psychiatry* 172 (10): 967–77.

doi:10.1176/appi.ajp.2015.14101266.

Nestor, Sean M, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and the Alzheimer's Disease Neuroimaging Alzheimer's Disease Neuroimaging Initiative. 2008. "Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated Using the Alzheimer's Disease Neuroimaging Initiative Database." *Brain : A Journal of Neurology* 131 (Pt 9). Oxford University Press: 2443–54. doi:10.1093/brain/awn146.

Noordermeer, Siri D. S., Marjolein Luman, and Jaap Oosterlaan. 2016. "A Systematic Review and Meta-Analysis of Neuroimaging in Oppositional Defiant Disorder (ODD) and Conduct Disorder (CD) Taking Attention-Deficit Hyperactivity Disorder (ADHD) Into Account." *Neuropsychology Review* 26 (1): 44–72. doi:10.1007/s11065-015-9315-8.

Noordermeer, Siri D S, Marjolein Luman, Corina U Greven, Kim Veroude, Stephen V Faraone, Catharina A Hartman, Pieter J Hoekstra, et al. 2017. "Structural Brain Abnormalities of Attention-Deficit/Hyperactivity Disorder With Oppositional Defiant Disorder." *Biological Psychiatry* 82 (9). Elsevier: 642–50. doi:10.1016/j.biopsych.2017.07.008.

Norman, Luke J., Christina Carlisi, Steve Lukito, Heledd Hart, David Mataix-Cols, Joaquim Radua, and Katya Rubia. 2016. "Structural and Functional Brain Abnormalities in Attention-Deficit/Hyperactivity Disorder and Obsessive-Compulsive Disorder." *JAMA Psychiatry* 73 (8): 815. doi:10.1001/jamapsychiatry.2016.0700.

Rommelse, Nanda N.J., Hilde M. Geurts, Barbara Franke, Jan K. Buitelaar, and Catharina A. Hartman. 2011. "A Review on Cognitive and Brain Endophenotypes That May Be Common in Autism Spectrum Disorder and Attention-Deficit/hyperactivity Disorder and Facilitate the Search for Pleiotropic Genes." *Neuroscience & Biobehavioral Reviews* 35 (6): 1363–96. doi:10.1016/j.neubiorev.2011.02.015.

Schweren, Lizanne J.S., Catharina A. Hartman, Dirk J. Heslenfeld, Dennis van der Meer, Barbara Franke, Jaap Oosterlaan, Jan K. Buitelaar, Stephen V. Faraone, and Pieter J. Hoekstra. 2015. "Thinner Medial Temporal Cortex in Adolescents With Attention-Deficit/Hyperactivity Disorder and the Effects of Stimulants." *Journal of the American Academy of Child & Adolescent Psychiatry* 54 (8): 660–67. doi:10.1016/j.jaac.2015.05.014.

Thapar, A, J Holmes, K Poulton, and R Harrington. 1999. "Genetic Basis of Attention Deficit and Hyperactivity." *The British Journal of Psychiatry : The Journal of Mental Science* 174 (February): 105–11. http://www.ncbi.nlm.nih.gov/pubmed/10211163.

Thomas, Rae, Sharon Sanders, Jenny Doust, Elaine Beller, and Paul Glasziou. 2015. "Prevalence of Attention-Deficit/hyperactivity Disorder: A Systematic Review and Meta-Analysis." *Pediatrics* 135 (4). American Academy of Pediatrics: e994-1001. doi:10.1542/peds.2014-3482.

Thompson, Paul M, Kiralee M Hayashi, Greig I De Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, Michael S Hong, et al. 2004. "Mapping Hippocampal and Ventricular Change in Alzheimer Disease." doi:10.1016/j.neuroimage.2004.03.040.

Torres, Imma, Marina Garriga, Brisa Sole, Caterina M. Bonnín, Montse Corrales, Esther

Jiménez, Eva Sole, et al. 2017. “Functional Impairment in Adult Bipolar Disorder with ADHD.” *Journal of Affective Disorders* 227 (February): 117–25. doi:10.1016/j.jad.2017.09.037.

Willcutt, Erik G. 2012. “The Prevalence of DSM-IV Attention-Deficit/Hyperactivity Disorder: A Meta-Analytic Review.” *Neurotherapeutics* 9 (3): 490–99. doi:10.1007/s13311-012-0135-8.

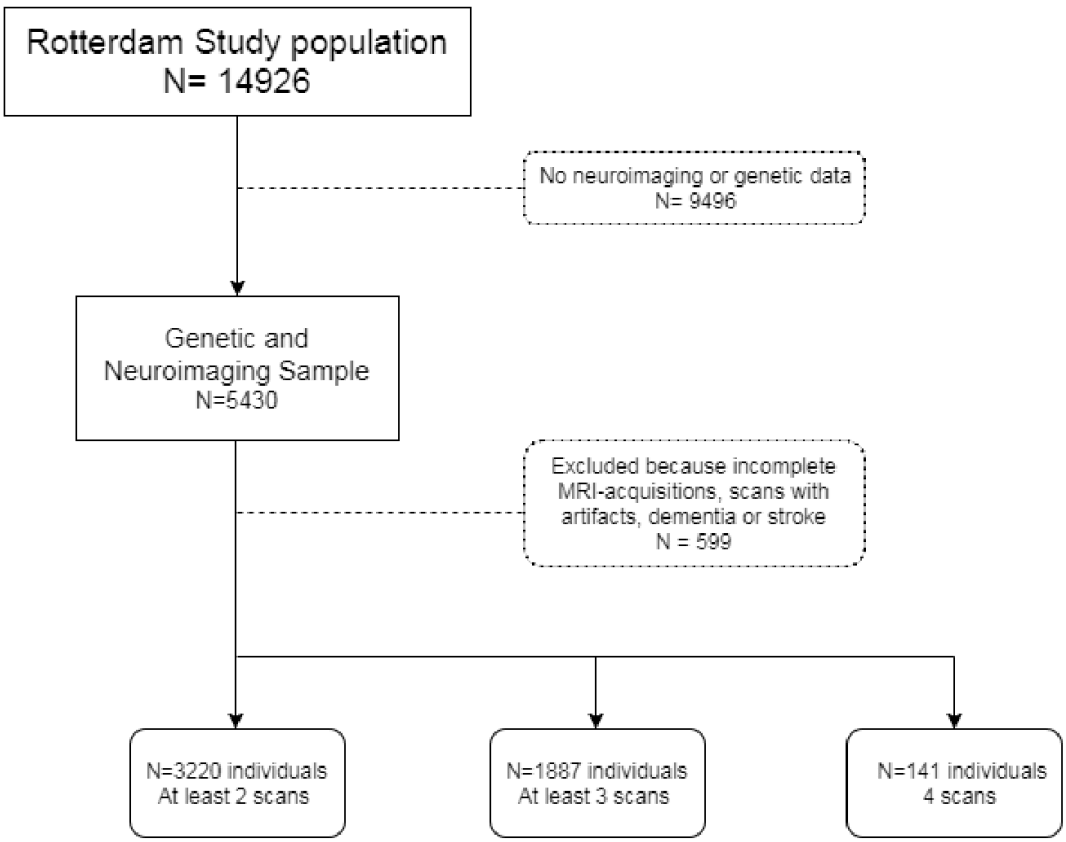

Figure 1. Flowchart of the sample of study

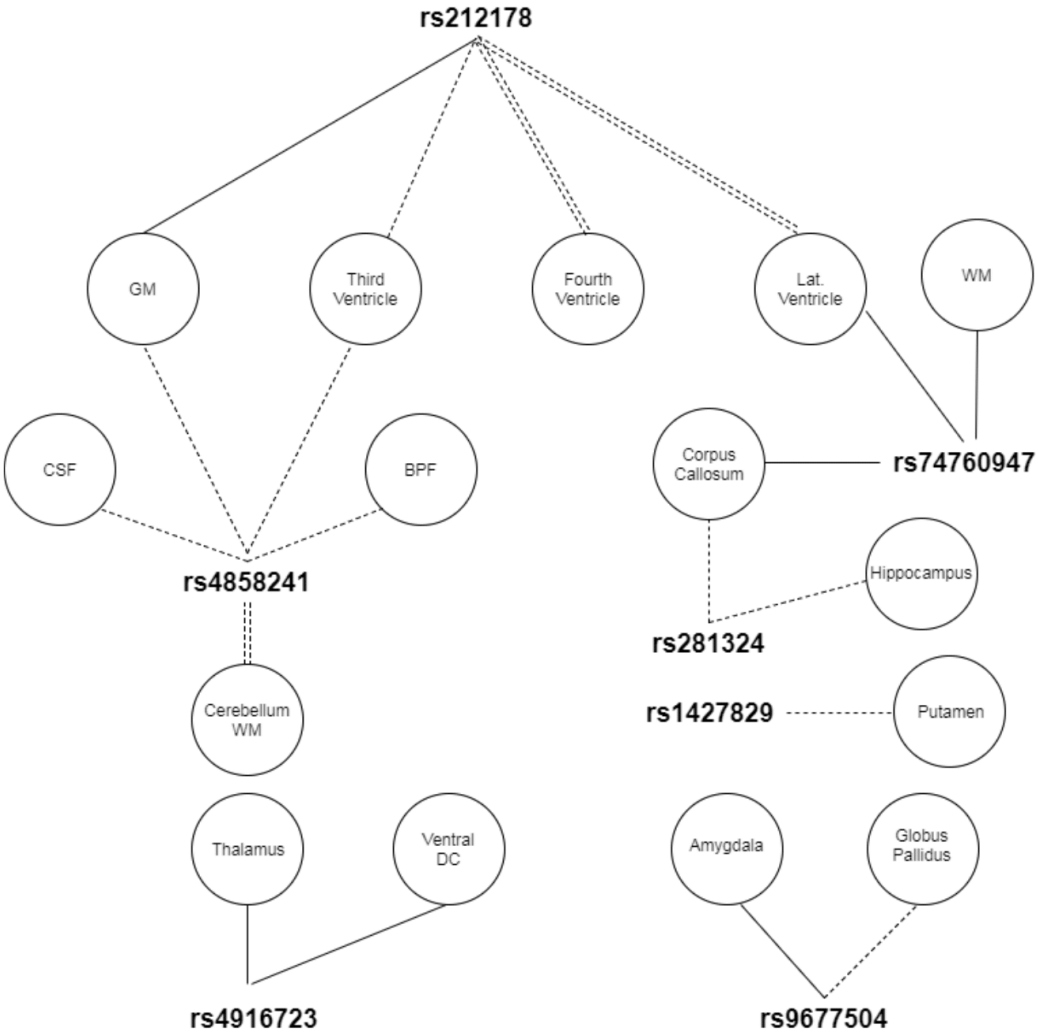

Figure 2. Cross sectional design. Age-by-SNP results. Straight lines represents a significant age-by-SNP interaction at the first spline. Dotted lines represents a significant age-by-SNP interaction at the second spline. Parallel dotted lines represents a significant age-by-SNP interaction term in both age splines.

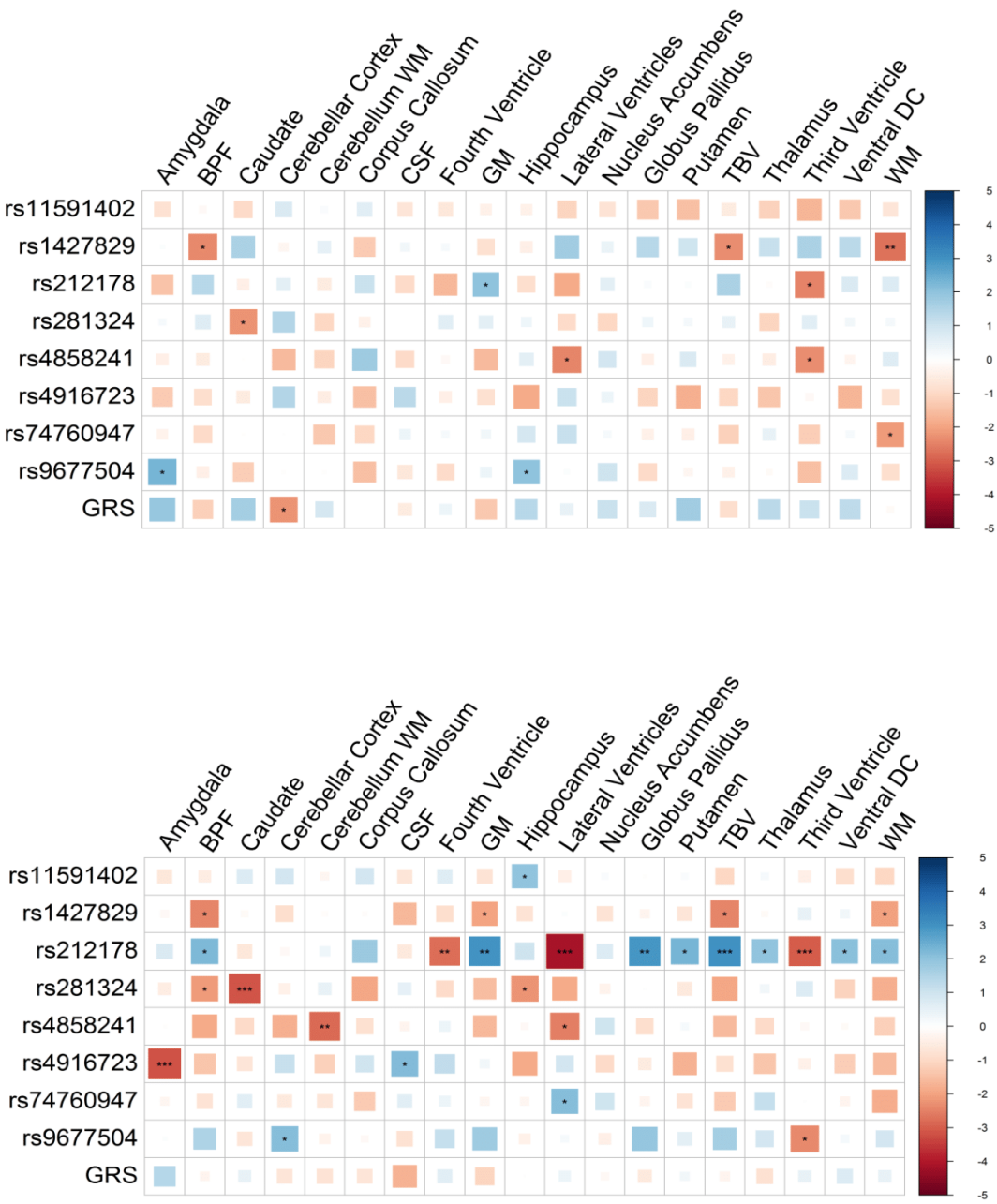

Figure 3. Longitudinal design. Age-by-SNP results. ***Pvalue =0.003; **Pvalue = 0.01; *Pvalue = 0.05.

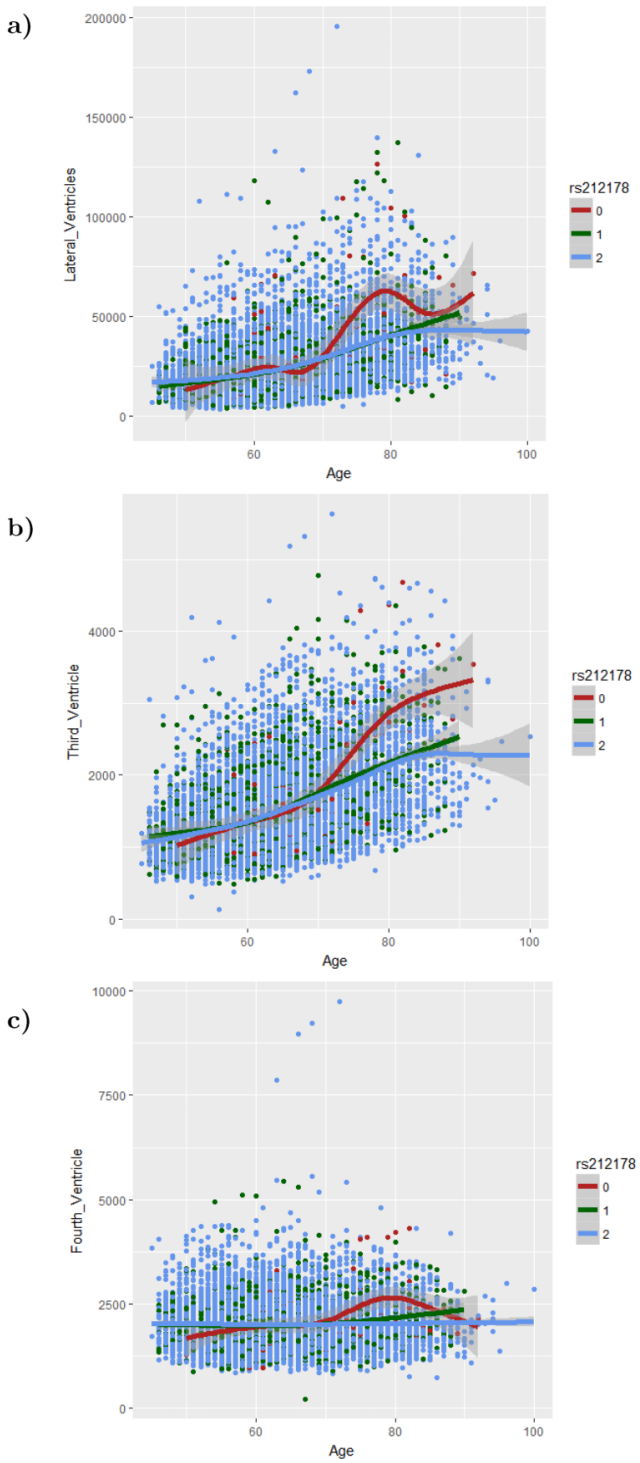

Figure 4. Trajectory differences of ventricle structures with age between rs212178 genotypes. (**a**) lateral ventricles; (b) third ventricle; (c) fourth ventricle.

	N	Age in years, mean (SD), range	Sex distribution (F/M)(%)
scan 1	4831	64.7(10.9), 45.0-97.0	2659/2172 (55.4%)
scan 2	3220	65.3(9.3), 50-100	1731/1489 (53.8%)
scan 3	1887	64.5(7.3), 51-94	989/898 (52.4%)
scan 4	141	74.8(3.9), 55-89	59/82 (41.8%)

Table 1. Participant characteristics.

SNP	CHR	position	A1	A2	MAF	Gene
rs9677504	2	215.181.889	G	A	0,09	*SPAG16*
rs4858241	3	20.669.071	T	G	0,39	*Intergenic*
rs4916723	5	87.854.395	A	C	0,44	*LINC00461, MIR9-2, LINC02060*
rs74760947	8	34.352.610	A	G	0,04	*LINC01288*
rs11591402	10	106.747.354	T	A	0,22	*SORCS3*
rs1427829	12	89.760.744	A	G	0,44	*DUSP6, POC1B*
rs281324	15	47.754.018	C	T	0,47	*SEMA6D*
rs212178	16	72.578.131	A	G	0,10	*LINC01572*

Table 2. Full summary statistics for all SNPs tested.

Table 3. Descriptives of brain structures.

	scan 1; N=4,831				*scan 2; N=3,220*			
Variable	**Mean(SD)**	**Min**	**Median**	**Max**	**Mean(SD)**	**Min**	**Median**	**Max**
Amygdala, mm3	2,713 (393)	1.045	2.709	4.448	2750.6(389.04)	1.032	2.749	4.857
Caudate Nucleus, mm3	6,896 (1,099)	3.454	6.745	14.912	6859.1(1077.93)	4.453	6.724	13.598
Cerebellum Cortex, mm3	99,129(11,068)	34.039	98.882	148.289	99918.4(11039.07)	58.138	99.575	143.131
Cerebellum WM, mm3	23,679 (3,517)	6.398	23.476	44.160	23759.2(3466.06)	13.135	23.502	49.391
Corpus callosum, mm3	2,625 (548)	370	2.652	4.601	2628.2(541.04)	643	2.643	4.521
Fourth Ventricle, mm3	1,996 (576)	218	1.907	7.865	2035.9(590.22)	734	1.954	8.967
Globus Pallidus, mm3	2,895 (471)	1.377	2.870	6.318	2886.1(449.35)	1.427	2.876	5.367
Hippocampus, mm3	7,707 (1,079)	2.191	7.793	11.722	7833.1(1071.05)	3.117	7.932	11.277
Lateral Ventricles, mm3	26,464 (16,708)	2.900	21.832	146.082	26785(16869.89)	3.908	21.996	162.490
Nucleus Accumbens, mm3	1,049 (176)	264	1.042	2.159	1051.5(173.97)	507	1.042	1.885
Putamen, mm3	9,102 (1,280)	3.965	9.019	17.775	9103.1(1230.39)	5.123	9.061	20.864
Thalamus, mm3	12,554 (1,567)	7.178	12.443	26.764	12562.8(1524.85)	8.175	12.490	29.273
Third Ventricle, mm3	1,551 (656)	308	1.403	4.822	1580.5(660.05)	128	1.430	5.189
Ventral DC, mm3	7679(884)	4.881	7.631	11.744	7777(881)	4.949	7.750	11.771
ICV, mm3	1,479,696 (161,628)	926.412	1.475.734	2.075.800	1486431.8(161509.86)	961.537	1.483.578	2.105.942
GM, mm3	596,235 (58,456)	396.242	593.973	802.581	598063.5(57164.82)	406.333	595.949	795.062
WM, mm3	434,979 (64,998)	204.521	434.070	717.608	435958.6(62887.63)	190.808	435.316	656.974
TBV, mm3	1,031,215 (118,976)	640.449	1.028.389	1.510.603	1034022.1(115598.59)	684.065	1.030.530	1.434.771
BPF, %	0.7 (0.05)	0	1	1	0.7(0.05)	0	1	1
CSF, mm3	1,266 (367)	482	1.196	4.303	1286.7(356.69)	571	1.218	3.819

Table 3. Descriptives of brain structures (continue).

	scan 3; N=1,887				*scan 4; N=141*			
Variable	**Mean(SD)**	**Min**	**Median**	**Max**	**Mean(SD)**	**Min**	**Median**	**Max**
Amygdala, mm3	2759.1(374.07)	1492	2751	4232	2602.6(417.55)	1.384	2.575	3.645
Caudate Nucleus, mm3	6805.4(1018.54)	4299	6667	11463	6875.5(1221.42)	4.901	6.660	12.620
Cerebellum Cortex, mm3	100763.9(10792.79)	70256	99930	138059	96941(10020.02)	73.563	97.483	121.866
Cerebellum WM, mm3	23978.7(3446.2)	12854	23695	41765	21993(2842.77)	13.123	21.882	30.297
Corpus callosum, mm3	2651(521.21)	624	2653	4882	2343.6(557.24)	747	2.383	3.811
Fourth Ventricle, mm3	2042.5(592.75)	842	1964	9231	2121.9(866.68)	932	2.006	9.732
Globus Pallidus, mm3	2943.9(457.94)	1629	2922	4974	2702.1(384.8)	1.813	2.694	3.790
Hippocampus, mm3	7922.4(1040.71)	3731	7994	11218	7237.3(1034.33)	4.190	7.359	9.758
Lateral Ventricles, mm3	25662.2(15796.38)	4623	21580	172963	36200.9(21934.68)	9.002	30.541	195.491
Nucleus Accumbens, mm3	1059.3(173.65)	391	1049	1705	1005.3(164.91)	651	998	1.640
Putamen, mm3	9158.7(1220.54)	5274	9064	15568	8621.4(1097.71)	6.377	8.443	11.104
Thalamus, mm3	12685.8(1542.15)	7999	12625	33813	11851(1306.25)	8.767	11.750	16.874
Third Ventricle, mm3	1533(597.54)	376	1396	5320	1977.8(744.44)	814	1.835	5.629
Ventral DC, mm3	7854(873)	5092	7818	13968	7429(767)	5.494	7.453	9.543
ICV, mm3	1491225.6(158877.95)	991952	1483989	2104746	1468569.5(174478.89)	1.009.685	1.463.542	1.946.649
GM, mm3	601487.7(55335.99)	414281	599142	782188	574293.4(49750.73)	445.626	572.779	717.230
WM, mm3	439350.2(58896.54)	238022	438587	646887	406666.3(58138.18)	262.422	404.217	593.625
TBV, mm3	1040837.9(109658.07)	652303	1037746	1388786	980959.7(102646.47)	745.082	974.872	1.282.747
BPF, %	0.7(0.04)	0	1	1	0.7(0.04)	1	1	1
CSF, mm3	1287.2(351.46)	583	1221	4911	1431.2(451.56)	766	1.341	4.569

	GM, mm3		WM, mm3		TBV, mm3		BPF, %		CSF, mm3	
	β (95%CI)	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*
rs11591402	632.3(-883,2147.5)	0,414	1290.1(-634.1,3214.3)	0,189	1922.4(-953.6,4798.4)	0,19	0.1(-0.1,0.3)	0,336	-15.8(-32.5,0.9)	0,064
rs1427829	180.8(-1068.5,1430)	0,777	889.6(-696.8,2475.9)	0,272	1070.3(-1300.9,3441.5)	0,376	0.1(-0.1,0.2)	0,327	-8.6(-22.4,5.2)	0,22
rs212178	-2809(-4788.1,-829.9)	**0,005**	-693.1(-3209.6,1823.5)	0,589	-3502.1(-7261.6,257.4)	0,068	-0.3(-0.5,0)	**0,049**	18.2(-3.7,40.1)	0,103
rs281324	760.4(-496.5,2017.3)	0,236	1088(-508.3,2684.3)	0,182	1848.4(-537.3,4234.1)	0,129	0.1(0,0.3)	0,165	-6.3(-20.2,7.6)	0,376
rs4858241	-331.7(-1588,924.7)	0,605	-95.4(-1691,1500.3)	0,907	-427(-2811.9,1957.9)	0,726	0(-0.2,0.1)	0,604	-1.1(-14.9,12.8)	0,881
rs4916723	-537.5(-1776.1,701.1)	0,395	-751.8(-2324.8,821.2)	0,349	-1289.3(-3640.3,1061.7)	0,283	-0.1(-0.3,0.1)	0,228	2.5(-11.2,16.2)	0,718
rs74760947	-122.5(-3345.8,3100.8)	0,941	-988.8(-5082.5,3104.8)	0,636	-1111.3(-7229.9,5007.3)	0,722	-0.1(-0.5,0.3)	0,737	17.9(-17.7,53.6)	0,323
rs9677504	15.9(-2149,2180.7)	0,989	1826.5(-922.3,4575.3)	0,193	1842.4(-2266.7,5951.4)	0,38	0.1(-0.1,0.4)	0,357	-7.5(-31.5,16.4)	0,537
GRS	1503(-4403,7409.1)	0,618	2010.9(-5490.1,9511.9)	0,599	3513.9(-7697.3,14725.1)	0,539	0.3(-0.4,1.1)	0,388	-25(-90.3,40.3)	0,453
	Hippocampus, mm3		**Amygdala, mm3**		**Thalamus, mm3**		**Ventral DC, mm3**		**Corpus callosum, mm3**	
	β (95%CI)	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*
rs11591402	-14(-61,33)	0,56	5.9(-11.1,22.9)	0,495	-11.4(-69.9,47)	0,701	8.5(-25.9,42.9)	0,627	34.5(7.5,61.4)	**0,012**
rs1427829	-18.2(-56.9,20.5)	0,357	-0.1(-14.1,13.9)	0,988	-9.2(-57.4,39)	0,708	-6.7(-35.21.7)	0,646	20(-2.2,42.2)	0,078
rs212178	-56.9(-118.3,4.5)	0,07	-15.9(-38.2,6.3)	0,16	-81.3(-157.7-4.9)	**0,037**	-40.4(-85.3,4.6)	0,079	-7(-42.2,28.3)	0,699
rs281324	6.5(-32.5,45.5)	0,745	-2.4(-16.6,11.7)	0,734	5.3(-43.2,53.8)	0,831	9.3(-19.3,37.8)	0,524	27.3(5,49.7)	**0,017**
rs4858241	-15.7(-54.7,23.2)	0,429	-5.3(-19.4,8.9)	0,466	-5(-53.5,43.5)	0,839	10.5(-18,39)	0,471	-10.4(-32.7,11.9)	0,361
rs4916723	4.6(-33.8,43)	0,815	-8(-21.9,5.9)	0,258	15.4(-32.4,63.2)	0,529	15.7(-12.4,43.8)	0,273	5.8(-16.3,27.8)	0,609
rs74760947	-42(-141.9,58)	0,411	-17.6(-53.7,18.6)	0,342	-69.3(-193.6,55.1)	0,275	-26.1(-99.2,47.1)	0,485	-43(-100.3,14.3)	0,141
rs9677504	-64.8(-131.9,2.3)	0,058	-12.3(-36.7,12)	0,319	7.7(-75.8,91.3)	0,856	2.6(-46.6,51.7)	0,919	8.1(-30.4,46.6)	0,681
GRS	-31.4(-214.5,151.8)	0,737	16.6(-49.7,82.9)	0,624	87.4(-140.5,315.3)	0,452	11(-123.1,145)	0,872	-46.2(-151.3,58.8)	0,388

Table 4. Cross-sectional association results.

	Lateral Ventricles, mm3		Third Ventricle, mm3		Fourth Ventricle, mm3		Cerebellum WM, mm3		Cerebellum Cortex, mm3	
	β (95%CI)	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*
rs11591402	-332.2(-1040.6,376.1)	0,358	-10(-36.9,16.9)	0,467	15.4(-17.2,48.1)	0,355	-102(-263,59)	0,215	106.3(-385.9,598.5)	0,672
rs1427829	43.9(-540.1,627.9)	0,883	-7.7(-29.9,14.4)	0,494	-28.9(-55.8,-2)	**0,035**	16.1(-116.6,148.9)	0,812	26.2(-379.5,431.9)	0,899
rs212178	322.4(-603.8,1248.7)	0,495	-12.3(-47.5,22.9)	0,493	10.9(-31.8,53.6)	0,616	-152.2(-362.7,58.3)	0,156	-644.2(-1287.4,-1.1)	**0,05**
rs281324	-497.9(-1085.4,89.5)	0,097	-7.2(-29.5,15.1)	0,525	11.5(-15.5,38.6)	0,404	79.4(-54.2,212.9)	0,244	340.8(-67.3,748.9)	0,102
rs4858241	-226.5(-813.7,360.8)	0,45	-5.8(-28.1,16.5)	0,61	30.5(3.4,57.5)	**0,027**	5.6(-127.9,139.1)	0,934	-208(-615.9,200)	0,318
rs4916723	59.1(-520,638.2)	0,841	0.2(-21.8,22.2)	0,984	13.8(-12.9,40.5)	0,31	-36.1(-167.8,95.5)	0,59	-163.6(-565.8,238.7)	0,426
rs74760947	360.4(-1146.4,1867.2)	0,639	28.8(-28.4,86)	0,324	33.2(-36.3,102.6)	0,349	54.1(-288.4,396.6)	0,757	477.4(-569.3,1524.1)	0,371
rs9677504	-157.1(-1169.1,855)	0,761	-13(-51.4,25.5)	0,508	-17.4(-64.29.2)	0,465	227.3(-2.6,457.2)	0,053	189.7(-513.4,892.7)	0,597
GRS	186.6(-2574.5,2947.7)	0,895	-14.3(-119.2,90.5)	0,789	-109.9(-237.1,17.3)	0,09	436.3(-191.1,1063.7)	0,173	-205.6(-2123.8,1712.6)	0,834
	Caudate Nucleus, mm3		**Globus Pallidus, mm3**		**Putamen, mm3**		**Nucleus Accumbens, mm3**			
	β (95%CI)	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*	*β (95%CI)*	*P-value*		
rs11591402	-9.2(-60.1,41.7)	0,723	16(-5.4,37.5)	0,143	30.1(-27,87.2)	0,301	7.1(-1.7,16)	0,115		
rs1427829	-7.4(-49.3,34.5)	0,73	-12.2(-30,5.5)	0,176	-12.8(-59.9,34.2)	0,593	0.6(-6.7,7.9)	0,871		
rs212178	-27.7(-94.3,38.8)	0,414	-13.8(-41.9,14.3)	0,336	-68.7(-143.3,5.9)	0,071	-3.9(-15.5,7.7)	0,506		
rs281324	1.8(-40.4,44)	0,934	3.5(-14.3,21.3)	0,7	9.6(-37.8,56.9)	0,692	-0.9(-8.2,6.5)	0,818		
rs4858241	-39.9(-82.1,2.2)	0,063	-14.8(-32.6,3)	0,103	-47.4(-94.7,-0.1)	**0,049**	-0.9(-8.3,6.4)	0,808		
rs4916723	8.9(-32.7,50.5)	0,675	4.2(-13.3,21.8)	0,637	-16.6(-63.3,30)	0,485	2.5(-4.8,9.7)	0,501		
rs74760947	-70.3(-178.5,37.9)	0,203	-8.2(-53.9,37.5)	0,725	-111.3(-232.6,10)	0,072	3.1(-15.8,21.9)	0,751		
rs9677504	-7.2(-79.9,65.5)	0,847	-0.4(-31.1,30.3)	0,978	-1.2(-82.7,80.3)	0,977	2.4(-10.2,15.1)	0,707		
GRS	-23.5(-221.8,174.9)	0,817	-64.7(-148.5,19)	0,13	2.1(-220.3,224.5)	0,985	-9.2(-43.8,25.3)	0,601		

Table 4. Cross-sectional association results (continue).

Chapter 7

General Discussion

This thesis was aimed at developing and standardising statistical strategies to improve the assessment of potential relationships between genes and brain structures associated with neurodevelopmental domains. The derived results represent both methodological and neuropsychological contributions. From a methodological perspective, novel algorithms were developed in order to create a better model to fit and integrate imaging and genetic data. From a neuropsychological perspective, several genetic variants were demonstrated to be associated with ADHD symptoms and attention function. Furthermore, neuroimaging correlates of some of those variants and ADHD symptoms were explored. The results of the different studies included in the present thesis were presented and discussed in the previous results sections. Here, a more general discussion is provided.

7.1 Thesis contributions

The analysis of genetic and neuroimaging data is extremely relevant to gaining insight into the underlying biology of neurodevelopmental domains[71–73]. While progress has been made in both fields, these two sources of data have been traditionally studied in a parallel or independent manner. The joint analysis of imaging and genetic data can be an even more powerful mechanism by which to reveal biological mechanisms that would otherwise remain hidden. The study of human genetic variation has advanced from genotyping candidate regions to genome-wide genotyping and whole genome sequencing. Likewise, functional and structural brain maps have been created by studying common patterns of activation and deactivation, and common morphological areas of the brain. The integration of brain maps with genetic maps represents an excellent opportunity to deepen neuroscientific

knowledge. The main challenge in the Imaging Genetics (IG) field is to reduce complexity while capturing as much meaningful information in the data as possible. Different strategies and a variety of computational tools have been developed, but no established analytical framework exists as yet. As a result, the IG field remains challenging (see ***chapter 2***).

One extended analytical strategy is the inclusion of the relationship between the genetic components showing suggestive evidence of association with the phenotype of interest with neuroimaging phenotypes in a two-step procedure (see ***chapter 3***). In the first step, a subset of significant SNPs associated with neurodevelopmental domains is selected. In the second step, those SNPs associated with a specific neurodevelopmental domain are tested for association with brain structure and function. In ***chapter 4 section 1***, we considered the use of this strategy on attention function domains, detecting 13 potentially significant loci in the first step. One of these loci, rs4321351 (intronic at *PID1* gene[74]), was nominally replicated in an independent sample. In the second step, to further understand the role of this locus of potential interest for cognition, we examined its marginal effects on brain structure, function, and connectivity. These analyses revealed significant associations with frontal-basal ganglia circuits, suggesting the importance of basal ganglia in forming a complex functional system implicated in sustained attention processes. These results, together with those obtained in neuroimaging analysis, suggest for the first time the possibility that this SNP may play a role in the neuronal structure and functioning related to attention function domains. It is worth mentioning that most of the relevant findings involved reaction time variability –one of the most

replicated deficits in ADHD[75]. Previous research has highlighted reaction time as a promising cognitive target for molecular genetics investigation of these symptoms[76–78]. Moreover, evidence from previous studies also suggests that the cognitive functions assessed in nondemented populations may share common genetic factors with neurodegenerative disorders[79].

However, although the GWAS strategy is well known and the number of reported associations with health outcomes is extensive[80], the small effect sizes of variants, dimensionality of the data, and therefore the requirement for stringent statistics because of multiple testing, make it challenging to perform GWAS powerful enough to successfully map target SNPs/genes for neurodevelopmental domains.

Dimensionality of the data

The dimensionality of the data in IG studies represents the principal handicap in the acquisition of analytical perspectives on neurodevelopmental domains and complex neurological diseases, amplifying existing issues of reliability and interpretability of the results. Moreover, independently testing millions of SNPs population-wide for associations on the one hand, and millions of neuroimaging-based measures on the other, reduces statistical power due to multiple testing correction. Because statistical power is critical, since the most interesting effects are usually small and on the edge of detection, this creates controversy.

Multiple testing

Large-scale analysis of genetic vs. brain data requires new strategies for controlling Type-I error (incorrect rejection of a

true null hypothesis, also known as a false positive finding). Although most of the research context generally accepts a rate of confidence of 95 per cent, the hundreds of thousands of statistical tests composing an IG analysis increases the likelihood of an event and, therefore, of incorrectly rejecting the null hypothesis (raising the possibility of a Type I error). Hence, a rate of confidence of 95 per cent seems insufficient to properly account for multiple testing in IG studies. The most widely known and used method designed to counteract this problem, especially in genetics, is the Bonferroni correction. The Bonferroni correction defines a new confidence rate taking into account the number of tests performed in the analysis. However, this method has also received criticism for being very conservative[81], particularly in situations where there are a large number of tests and/or the test statistics are correlated (in linkage disequilibrium (LD), in the genetic context). In this situation, use of the Bonferroni correction comes at the cost of increasing the probability of producing false negatives, and in turn, decreases the power of the test (i.e., the probability of correctly rejecting a false null hypothesis). For this reason, in the genetic context, thresholds of significance have been established by agreement, based on a simulation study taking into account LD between SNPs. Therefore, genome-wide significance was set at $\alpha = 5x10^{-8}$, and suggestive evidence of association was set at $\alpha = 10^{-5}$. The question of choosing an appropriate threshold becomes murkier in the neuroimaging context. In the genetic context, the Bonferroni correction is an extended option available to correct for multiple testing. However, the significant spatial correlation of the neuroimaging data, for instance, in the simultaneous study of grey matter volume and subcortical structures, or cortical thickness and cortical surface areas, means the method is not optimal in this

context either. Alternatives include the use of Gaussian Random Field Theory[82], and/or non-parametric permutation correction techniques[83], which emerged as an ideal choice of number of effective tests to allow for adequate correction while maintaining high sensitivity.

Gene-set analysis

The gene-set analysis (GSA) methods are now in extensive use to reduce the number of statistical tests or dimensionality, and to increase the power of the hypothesis-free GWAS. GSA combine the effects of a set of single-variant signals and functional annotations within appropriate functional units (such as pathways or networks), which therefore requires fewer comparisons, increasing the statistical power of the analysis to detect significant signals and improving the interpretability of the results (see ***chapter 5 sections 1 and 2***). These are precisely the two main goals that GSA methods aims to combine: an understanding of the molecular mechanisms underlying neurodevelopmental domains, and the discovery of target variants to improve disease treatment. This improves GWAS performance because the functions and biological meanings of a set of variants passing a pre-defined p-value threshold in association tests cannot be inferred from p-values alone. On the other hand, the methods of analysis of the most popular sets of genes are based on empirical p-values that require a large number of permutations, which produces serious computational limitations. The proposed globalEVT method dramatically reduces the computational limitations of permutational procedures by considering a semiparametrical distribution for obtaining gene-level p-values. Therefore, while permutational GSA procedures do not allow for application to large data

sets of variants[84], globalEVT can be applied – without any computational limitations – in the large data sets that we find in practice.

Feature selection and dimensionality reduction

The extraction of interesting genetic and imaging-based features from complex IG studies often relies on feature selection. The use of feature selection methods can also be seen as an alternative strategy of GWAS for the identification of informative genetic markers related to neurodevelopmental domains, reducing the dimensionality of the data. The main benefits of this process are twofold: first, it removes any redundant features, which may improve prediction accuracy in addition to supporting the interpretability of results. Second, it helps in the generation of post-hoc inferences. In ***chapter 4 section 3***, we proposed the application of a two-step analysis combining a feature selection strategy designed for a count data distribution with a dimensionality reduction approach. This proposed strategy was inspired by the conclusions of earlier research, which indicated that the use of a feature selection strategy specifically for count data outcomes is more effective at reducing the dimensionality of the data.

On the other hand, decomposing the data into its most important sources of variation using dimensionality reduction methods holds the potential to discover unanticipated sources of signals with biological meaning, and generate new hypotheses (see ***chapter 5 section 3***). This strategy takes into account the structure of the genetic data and imaging markers and reduces the computational burden posed by large amounts of data. Furthermore, most variants identified confer

relatively small risk increments, and the amount of variability that is explained by these genetic components, expected for traits showing a polygenic architecture, is relatively small[85]. The proposed strategy, together with those already described in ***section 5***, makes it possible to determine the degree to which the whole set of genetic and/or neuroimaging markers contribute to the variability of the symptomatology jointly, rather than individually.

Modelling questionnaire-symptoms distribution

Neurodevelopmental research typically uses a dichotomous status – one that indicates whether the disorder is present or absent – to describe neurodevelopmental domains. However, such dichotomous classification does not capture variability in populations, since individuals who present extreme score values are grouped with individuals whose symptoms are just above the diagnosis threshold. This dichotomous approach contrasts with the classification of symptoms along a continuum, or a dimensional spectrum, ranging from normal to dysfunctional. In this context, quantitative biomarkers can classify individuals using a continuous spectrum rather than dichotomous diagnoses. In addition, quantitative biomarkers take into account residual variation within groups of individuals that are classified as case-control diagnosis, thereby also capturing differences in disease severity. Such information makes continuous phenotypes statistically more powerful for detecting (genetic) effects. For instance, ADHD symptoms often present skewed and overdispersed distributions, and, to date, little attention has been paid to their efficient statistical modelling. In ***chapter 4 section 2***, we aimed to highlight the importance of investigating the most appropriate model for count-symptoms extracted from

questionnaires, instead of directly assuming a dichotomous status. While no associations were found when using a dichotomous definition, results modelling the symptoms through a count-data distribution revealed a genetic variant, rs273342 (in an intron of the *MAPRE* gene[86], associated with ADHD, which in turn was associated with perivascular volumes.

Brain intermediate phenotypes

Although pioneering research in IG studies has attempted to link functional polymorphisms directly to behaviour, reported findings have been weak and inconsistent. A possible reason therefor may be the considerable inter-individual differences in observed behavioural dimensions, as well as subjectivity in behavioural measures. In addition, and also remarkably, gene effects at the level of behaviour are mediated by effects on molecular configuration, which may bias information processing in brain circuits, ultimately mediating behavioural responses to environmental changes. In this context, brain imaging is increasingly being recognised as an intermediate phenotype enabling researchers to understand the complex association between genetics and behavioural or clinical phenotypes[16,87]. The assumption is that neuroimaging techniques are able to identify neural pathways through which genetic polymorphisms contribute to the emergence of variability in behaviour, and consequently identify the risk of psychopathology. Moreover, the use of intermediate imaging phenotypes allows for the capture of associations with neurodevelopmental domains by incorporating the correlation between genetic variants and neuroimaging features[88].

When modelling brain measurements, it is important to notice that, so far, few studies on neuroanatomical correlates

have exclusively focused on longitudinal changes influenced by genetic components. Longitudinal models allow us to estimate not only the mean values of a quantitative trait at the baseline, but also the rates of their changes into each group as a result of age, including a group-age interaction. Thus, compared to cross-sectional design, longitudinal design can provide increased statistical power by reducing the confounding effect of between-subject variability, and provide unique insights into the temporal dynamics of the underlying biological processes of neurodevelopmental domains[89].

In the IG context, meta-analyses have shown that changes in several global subcortical volumes in the human brain are influenced by gene effects[90]. However, few studies in IG have assessed genetic effects on longitudinal brain changes, even though structural changes that occur during development and ageing are related to mental health and general cognitive functioning. For instance, in ***chapter 6***, we observed significant differences in nonlinear trajectories in the ventricle structures (lateral ventricles and third ventricle) related to rs212178 genotypic variation. These trajectories indicate a decrease in the volume of these structures during the late adult lifespan and demonstrate the involvement of non-linear dynamics in parts of the brain that have been extensively discussed in the context of cognition[91,92], Alzheimer's disease[93,94], and Parkinson's disease[95].

7.2 Clinical perspectives

Our knowledge of the biological causes of and risk factors for ADHD and related neurodevelopmental domains is advancing; however, the key question is how to translate and implement these insights into improved prevention and intervention

strategies. An important consideration in this regard is that the evolution of neuroimaging techniques and genome-wide sequencing and genotyping methods, which can be rapidly and cost-effectively applied in clinical settings. The studies described in this thesis may not directly influence clinical practice in the short term, but some important points can nevertheless be taken into account. First, assessed risk factors are of potential interest to research on preventive practices and to make personalised medicine possible by targeting those risk factors, for example, by creating appropriate treatment or preventive strategies for people who are at increased genetic risk of developing a specific neurodevelopmental disease. Second, a related clinical implication is the use of MRI-based features in the prediction of neurodevelopmental domains. The clinical utility of MRI does not reside solely in the attainment of a personalised diagnosis but extends to the provision of personalised treatment. In this context, brain abnormalities detected at pre-symptomatic level can be used for both diagnosis and treatment. Third, the evolution of these new neuroimaging techniques and sequencing methods could translate into their use as fast and profitable tools in a clinical context by enabling the characterization – specifically the pathology – of the individual. Finally, relevant IG findings in relation to biological mechanisms (e.g., the identification of new target genes) might assist with the design of personalised disease-modifying drugs and treatments.

7.3 Thesis limitations and strengths

Several limitations must be considered when interpreting the present work. First, some of the studies discussed in this thesis are hypothesis-free analyses, for which the IG sample

size is modest. Second, data acquisition in imaging studies is highly sensitive to noise signals and batch effects produced during scanners' acquisition of data. This often results in real signal effects being obscured, complicating the discovery of significant signals. Third, most neurodevelopmental domains are reported by means of questionnaires. This is an important source of heterogeneity due to the subjective measurement and the types of instruments used to obtain these measures. For instance, neurodevelopmental information related to ADHD symptoms were reported by only one informant (teachers), which provides insuficient information results in a lack of information about the occurrence of these symptoms, and constitutes a subjective measure of the score of these symptoms. In addition, the questionnaires were completed only once, resulting in a lack of information about the evolution of the symptomatology. Fourth, in some studies we examined multiple variables under a massive univariate approach, which may increase the probability of Type 1 error. Moreover, we examined statistical methods that, in turn, provide greater complexity in the interpretation of the results. The optimal manner in which to interpret and visualise such results and the suitability of applied statistical techniques (e.g., multiple testing corrections) to such data-analysis require further exploration. Fifth, the studies discussed between ***chapter 4 section 3*** and ***chapter 6*** must be replicated in independent samples. Sixth, most of the methods are limited by the difficulty of linking genetic variants to specific genes, as it has been demonstrated that the gene closest to the SNP is not always the functional gene[96,97]. Moreover, our current knowledge of biological pathways is limited and hampered by dissimilarities subject to the differences between several databases. Finally, most of the studies included in this thesis were based on cross-sectional

designs. This is a critical issue relevant to the performance of the studies, due to the fact that cross-sectional designs can only capture the mean differences of a phenotype across the subgroups of subjects, instead of its trajectory, and commonly used regression models assume that the measurements are independent. Moreover, neuroimaging phenotypes can be subject to reverse causation with no utility for early detection of the disease. In this sense, studies that make use of multiple scans could increase the likelihood of genetic discoveries by using data from different time points.

Overall, this thesis offers several strengths in terms of advancing our understanding of the development of variability in brain structure through the use of genomic information, which may be related to complex neuropsychological diseases and domains and includes several features designed to overcome the limitations of previous studies, including: i) the use of the dimensional perspective on neurodevelopmental domains in population-based studies, which implies the inclusion of subclinical manifestations of the symptoms that are more widespread in the population than the clinical disorders; ii) the use of novel methods to integrate multimodal data, which allows for the combination of multiple phenotypes and covariates, controlling for Type I error, without reducing statistical power (Type II error); iii) the application of gene-set analyses, which helps increase the power of the study; iv) the inclusion of replication samples (INMA project) of a similar age and assessed with the same instruments; v) the incorporation of neuroimaging data to gain insight into the possible biological effects of certain genetic variants on ADHD; vi) further exploration using larger study samples (the Rotterdam Study); and vii) the inclusion of longitudinal repeated MRI-measures from a large

cohort with information about genetics and other important confounding factors, which provides a unique opportunity to test longitudinal brain changes related to genetic information.

7.4 Future research

The work presented in this thesis opens the way to new lines of investigation: i) the improvement of statistical power to identify factors underlying the reported associations, ii) the potential benefit gained through longitudinal assessment of brain modelling, iii) the combination of different types of neuroimaging modalities and omics data, and iv) the fostering of imaging gene-environmental interaction models.

Statistical power

Statistical power can be boosted through the use of larger samples, particularly by joining global consortia collaborations or by using publicly available data from biobanks. Nevertheless, statistical power can also be improved by reducing the measurement error in both fields. For instance, advances in DNA sequencing and MRI scans, and reductions in costs associated with these modalities, will ease the way for obtaining whole genome sequences and precise MRI-derived data. Moreover, we can increase statistical power in IG studies by using smarter data analysis and developing more powerful statistical techniques. More specifically, the amount of data collected for IG studies will benefit from the application of machine learning and deep learning techniques. These techniques make use of sophisticated training algorithms that improve both the acquisition of data and the power of computing parallel and distributed analyses.

Longitudinal designs

As described above, much of the existing work in the field is cross-sectional, which provides indeterminate information about changes in the structure and functionality of the brain over time. Gaining a better understanding of the longitudinal trajectories of brain development might potentially uncover the underlying biological characterisation of complex neurodevelopmental domains and diseases.

Combination of different types of modalities

A critical challenge in IG studies is to model even greater complexity of genetic effects on the brain. To date, most neuroimaging studies have examined a single pair of sources of biological data at a time (e.g., SNPs with structural MRI). For instance, the integration of genetic neuroimaging methods with epigenetics, known as imaging epigenetics, promises to provide deeper insights into the causative pathways though which genes and environment interact during life and impact human brain development (Lista et al. 2013; Hampton 2016). Other studies have also started to analyse proteomics and neuroimaging-based features as potential biomarkers of the basis for computing essential cell functions in order to identify the best proteomic model for the diagnosis, monitoring and prediction of complex neurological disorders[100,101].

Because the underlying neurobiology of psychopathology is likely highly complex, in the future the integration of multimodal types of data relevant to imaging genetics (e.g., genomics, transcriptomics, proteomics, metabolomics, morphological MRI, DTI, or functional MRI) will become essential, holding the potential to amplify the synergistic

value of imaging genetics[102–104]. In this context, some studies have demonstrated the integration of gene expression data and genetic markers in order to facilitate the detection of those genetic markers that are not only associated with brain intermediate phenotypes, but also highly expressed there[105]. Hence, there is promise in integrating multiple sources of neuroimaging with omics data, beyond just two modalities[106]. However, the integration of multiple data modalities represents yet another dimension that can be added on top of imaging and genetics, and may therefore also require the development of novel methods to facilitate such studies.

Imaging gene-environmental interaction (IGxE) models

Designs focused on the integration of imaging gene-environment interactions (IGxE) open up new sources of analysis by means of which to gain an understanding of the conditional mechanisms through which genes, environment and the brain interact to predict neurodevelopmental domains and risk of neuropsychological diseases[107]. Such designs represent an opportunity to not only integrate difference sources of omics and imaging features, but also to integrate target environmental sources relevant to the structure and functioning of the brain and, consequently, to identify neurodevelopmental domains and neuropsychological diseases.

Finally, more work is required to utilise genetics, neuroimaging and neurocognitive data to predict outcomes and treatment response[108], as has been accomplished in other areas of psychiatric research[109].

Chapter 8

Conclusions

The main conclusions of this thesis are set out below.

1. Despite the growing body of research on imaging genetics, there is no agreement in relation to the standardisation of statistical procedures for the joint analysis of genetic and neuroimaging data.

2. Two-step procedures constitute the most extensively used analytical strategy available to deal with the integrated analysis of genetic components, brain features and neurodevelopmental domains.

    a. The rs4321351 (*PID1*), which is associated with structural and functional brain changes in basal ganglia circuits, may be involved in attention function during childhood.

    b. The mTOR signalling and amyloid secretase precursor pathways, involved in the pathogenesis of Alzheimer's disease, may play a role in the development of attention function.

3. The research field of IG can benefit from the application of novel multivariate strategies that can better handle the size of genetic and imaging datasets and their inherent variability.

4. An appropriate modelling of neurodevelopmental domains based on a count of questionnaire symptoms can increase statistical power to establish significant risk factors.

    a. ADHD symptoms may be influenced by rs273342 genotypes, which in turn demonstrated a strong association with Virchow-Robin spaces if the

symptoms were modelled under a negative binomial distribution.

5. Gene-set analysis strategies allow for reductions in multiple testing and the incorporation of previous biological knowledge into the analysis.

6. The proposed gene-set analysis methods improve power by allowing different inheritance models for each genetic variant, and allow for the existence of correlation between genetic variants. Specifically, the globalEVT method improves computational efficiency, identifying genes which may play an important role in the mechanisms governing the development of complex diseases.

    a. Significant genes obtained from the application of gene-set developed methods encode a functional network configuration regulator of the ubiquitin-proteasome system, suggesting new biological mechanisms linked to ADHD.

7. Multivariate modelling facilitates the combined analysis of genetic and neuroimaging data, improving the understanding of the complex relationships affecting the multifactorial nature of neurodevelopmental domains and neurological diseases.

    a. White matter was the most important brain structure in terms of explained variability associated with ADHD symptoms and the inattention and hyperactivity domains.

    b. Executive functioning is mainly characterised by latent factors linked to changes in the brain

structure of grey matter, white matter, and the cerebellar and ventricular structures. In contrast, genetic characteristics have a reduced effect on the explained variability of the executive function domain.

8. A decrease in the volume of ventricular structures (lateral ventricles, and third ventricle) is related to rs212178 genotypic variation during the late adult lifespan.

REFERENCES

1. Alemany, S. *et al.* A Genome-Wide Association Study of Attention Function in a Population-Based Sample of Children. *PLoS One* **11,** e0163048 (2016).
2. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538,** 248–252 (2016).
3. Felix, J. F. *et al.* Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Hum. Mol. Genet.* **25,** 389–403 (2016).
4. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47,** 1449–56 (2015).
5. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
6. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (2009).
7. Ashburner, J. & Friston, K. J. Voxel-Based Morphometry—The Methods. *Neuroimage* **11,** 805–821 (2000).
8. Good, C. D. *et al.* A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. *Neuroimage* **14,** 21–36 (2001).
9. Lu, H. *et al.* Novel approach to the measurement of absolute cerebral blood volume using vascular-space-occupancy magnetic resonance imaging. *Magn. Reson. Med.* **54,** 1403–1411 (2005).
10. Ogawa, S. *et al.* Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model. *Biophys. J.* **64,** 803–812 (1993).
11. Gottesman, I. I. & Gould, T. D. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *Am. J. Psychiatry* **160,** 636–645 (2003).
12. Preston, G. A. & Weinberger, D. R. Intermediate phenotypes in schizophrenia: a selective review. *Dialogues Clin. Neurosci.* **7,** 165–79 (2005).
13. Braff, D. L. & Tamminga, C. A. Endophenotypes, Epigenetics, Polygenicity and More: Irv Gottesman's Dynamic Legacy. *Schizophr. Bull.* **43,** 10–16 (2017).
14. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric

disease. *Science* **349,** 1489–94 (2015).
15. Iacono, W. G., Vaidyanathan, U., Vrieze, S. I. & Malone, S. M. Knowns and unknowns for psychophysiological endophenotypes: Integration and response to commentaries. *Psychophysiology* **51,** 1339–1347 (2014).
16. Meyer-Lindenberg, A. & Weinberger, D. R. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nat. Rev. Neurosci.* **7,** 818–827 (2006).
17. Munafò, M. R. & Flint, J. The genetic architecture of psychophysiological phenotypes. *Psychophysiology* **51,** 1331–1332 (2014).
18. Larsson, H., Anckarsater, H., Råstam, M., Chang, Z. & Lichtenstein, P. Childhood attention-deficit hyperactivity disorder as an extreme of a continuous trait: a quantitative genetic study of 8,500 twin pairs. *J. Child Psychol. Psychiatry.* **53,** 73–80 (2012).
19. Lubke, G. H., Hudziak, J. J., Derks, E. M., van Bijsterveldt, T. C. E. M. & Boomsma, D. I. Maternal Ratings of Attention Problems in ADHD: Evidence for the Existence of a Continuum. *J. Am. Acad. Child Adolesc. Psychiatry* **48,** 1085–1093 (2009).
20. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.).* (2013).
21. Insel, T. *et al.* Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *Am. J. Psychiatry* **167,** 748–751 (2010).
22. MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. On the practice of dichotomization of quantitative variables. *Psychol. Methods* **7,** 19–40 (2002).
23. Marcus, D. K. & Barry, T. D. Does attention-deficit/hyperactivity disorder have a dimensional latent structure? A taxometric analysis. *J. Abnorm. Psychol.* **120,** 427–442 (2011).
24. Goodkind, M. *et al.* Identification of a Common Neurobiological Substrate for Mental Illness. *JAMA Psychiatry* **72,** 305 (2015).
25. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381,** 1371–1379 (2013).
26. Hudziak, J. J., Achenbach, T. M., Althoff, R. R. & Pine, D. S. A dimensional approach to developmental psychopathology. *Int. J. Methods Psychiatr. Res.* **16 Suppl 1,** S16-23 (2007).

27. Lubke, G. H., Hudziak, J. J., Derks, E. M., van Bijsterveldt, T. C. E. M. & Boomsma, D. I. Maternal ratings of attention problems in ADHD: evidence for the existence of a continuum. *J. Am. Acad. Child Adolesc. Psychiatry* **48,** 1085–93 (2009).
28. Polderman, T. J. C. *et al.* Across the continuum of attention skills: a twin study of the SWAN ADHD rating scale. *J. Child Psychol. Psychiatry* **48,** 1080–1087 (2007).
29. Willcutt, E. G. The Prevalence of DSM-IV Attention-Deficit/Hyperactivity Disorder: A Meta-Analytic Review. *Neurotherapeutics* **9,** 490–499 (2012).
30. Thomas, R., Sanders, S., Doust, J., Beller, E. & Glasziou, P. Prevalence of attention-deficit/hyperactivity disorder: a systematic review and meta-analysis. *Pediatrics* **135,** e994-1001 (2015).
31. Faraone, S. V. *et al.* Attention-deficit/hyperactivity disorder. *Nat. Rev. Dis. Prim.* **1,** 15020 (2015).
32. Franke, B. *et al.* The genetics of attention deficit/hyperactivity disorder in adults, a review. *Mol. Psychiatry* **17,** 960–87 (2012).
33. Faraone, S. V *et al.* Molecular genetics of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **57,** 1313–23 (2005).
34. Faraone, S. V & Mick, E. Molecular genetics of attention deficit hyperactivity disorder. *Psychiatr. Clin. North Am.* **33,** 159–80 (2010).
35. Larsson, H., Chang, Z., D'Onofrio, B. M. & Lichtenstein, P. The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan. *Psychol. Med.* **44,** 2223–9 (2014).
36. Bralten, J. *et al.* Candidate genetic pathways for attention-deficit/hyperactivity disorder (ADHD) show association to hyperactive/impulsive symptoms in children with ADHD. *J. Am. Acad. Child Adolesc. Psychiatry* **52,** 1204–1212.e1 (2013).
37. Banaschewski, T., Becker, K., Scherag, S., Franke, B. & Coghill, D. Molecular genetics of attention-deficit/hyperactivity disorder: an overview. *Eur. Child Adolesc. Psychiatry* **19,** 237–57 (2010).
38. Faraone, S. V, Doyle, A. E., Mick, E. & Biederman, J. Meta-analysis of the association between the 7-repeat allele of the dopamine D(4) receptor gene and attention deficit hyperactivity disorder. *Am. J. Psychiatry* **158,** 1052–7 (2001).
39. Maher, B. S., Marazita, M. L., Ferrell, R. E. & Vanyukov, M. M. Dopamine system genes and attention deficit hyperactivity disorder: a meta-analysis. *Psychiatr. Genet.* **12,** 207–15 (2002).

40. Demontis, D. *et al.* Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. *bioRxiv* 145581 (2017). doi:10.1101/145581
41. Bener, A., Kamal, M., Bener, H. & Bhugra, D. Higher prevalence of iron deficiency as strong predictor of attention deficit hyperactivity disorder in children. *Ann. Med. Health Sci. Res.* **4,** S291-7 (2014).
42. Doehnert, M., Brandeis, D., Schneider, G., Drechsler, R. & Steinhausen, H.-C. A neurophysiological marker of impaired preparation in an 11-year follow-up study of attention-deficit/hyperactivity disorder (ADHD). *J. Child Psychol. Psychiatry.* **54,** 260–70 (2013).
43. Hanć, T. *et al.* ADHD and overweight in boys: cross-sectional study with birth weight as a controlled factor. *Eur. Child Adolesc. Psychiatry* **24,** 41–53 (2015).
44. Kim, S. *et al.* Lead, mercury, and cadmium exposure and attention deficit hyperactivity disorder in children. *Environ. Res.* **126,** 105–10 (2013).
45. Ramos, R. *et al.* Association of ADHD symptoms and social competence with cognitive status in preschoolers. *Eur. Child Adolesc. Psychiatry* **22,** 153–164 (2012).
46. Frodl, T. & Skokauskas, N. Meta-analysis of structural MRI studies in children and adults with attention deficit hyperactivity disorder indicates treatment effects. *Acta Psychiatr. Scand.* **125,** 114–126 (2012).
47. Norman, L. J. *et al.* Structural and Functional Brain Abnormalities in Attention-Deficit/Hyperactivity Disorder and Obsessive-Compulsive Disorder. *JAMA Psychiatry* **73,** 815 (2016).
48. Noordermeer, S. D. S. *et al.* Structural Brain Abnormalities of Attention-Deficit/Hyperactivity Disorder With Oppositional Defiant Disorder. *Biol. Psychiatry* **82,** 642–650 (2017).
49. Schweren, L. J. S. *et al.* Thinner Medial Temporal Cortex in Adolescents With Attention-Deficit/Hyperactivity Disorder and the Effects of Stimulants. *J. Am. Acad. Child Adolesc. Psychiatry* **54,** 660–667 (2015).
50. Kumar, U., Arya, A. & Agarwal, V. Neural alterations in ADHD children as indicated by voxel-based cortical thickness and morphometry analysis. *Brain Dev.* **39,** 403–410 (2017).
51. Ambrosino, S., de Zeeuw, P., Wierenga, L. M., van Dijk, S. &

Durston, S. What can Cortical Development in Attention-Deficit/Hyperactivity Disorder Teach us About the Early Developmental Mechanisms Involved? *Cereb. Cortex* **27,** 4624–4634 (2017).

52. Polanczyk, G. & Rohde, L. A. Epidemiology of attention-deficit/hyperactivity disorder across the lifespan. *Curr. Opin. Psychiatry* **20,** 386–392 (2007).
53. Arcos-Burgos, M., Vélez, J. I., Solomon, B. D. & Muenke, M. A common genetic network underlies substance use disorders and disruptive or externalizing disorders. *Hum. Genet.* **131,** 917–929 (2012).
54. Durston, S. Imaging genetics in ADHD. *Neuroimage* (2010). Available at: http://www.ncbi.nlm.nih.gov/pubmed/20206707. (Accessed: 3rd September 2015)
55. Klein, M. *et al.* Brain imaging genetics in ADHD and beyond – Mapping pathways from gene to disorder at different levels of complexity. *Neurosci. Biobehav. Rev.* **80,** 115–155 (2017).
56. Posner, M. I. & Rothbart, M. K. Toward a physical basis of attention and self-regulation. *Phys. Life Rev.* **6,** 103–120 (2009).
57. Craig, F. *et al.* A review of executive function deficits in autism spectrum disorder and attention-deficit/hyperactivity disorder. *Neuropsychiatr. Dis. Treat.* **12,** 1191–202 (2016).
58. Franke, B. *et al.* The genetics of attention deficit/hyperactivity disorder in adults, a review. *Mol. Psychiatry* **17,** 960–87 (2012).
59. Barkley, R. A. The Important Role of Executive Functioning and Self-Regulation in ADHD©.
60. Bogdan, R. *et al.* Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential. *Biol. Psychiatry* **82,** 165–175 (2017).
61. Potkin, S. G. *et al.* A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophr. Bull.* **35,** 96–108 (2009).
62. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).
63. Stein, J. L. *et al.* Voxelwise genome-wide association study (vGWAS). *Neuroimage* **53,** 1160–74 (2010).
64. Shen, L. *et al.* Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI

and AD: A study of the ADNI cohort. *Neuroimage* **53,** 1051–63 (2010).

65. Hibar, D. P. *et al.* Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Mol. Psychiatry* (2017). doi:10.1038/mp.2017.73
66. Adams, H. H. H. *et al.* Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19,** 1569–1582 (2016).
67. Chi, E. C. *et al.* Imaging Genetics Via Sparse Canonical Correlation Analysis. *Proceedings. IEEE Int. Symp. Biomed. Imaging* **2013,** 740–743 (2013).
68. Vounou, M. *et al.* Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* **60,** 700–16 (2012).
69. Wan, J. *et al.* Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. *Med. Image Comput. Comput. Assist. Interv.* **14,** 376–83 (2011).
70. Vilor-Tejedor, N. GitHub Repository. (2017). Available at: https://github.com/natvt8.
71. Varcin, K. J., Nelson, C. A. & III. A developmental neuroscience approach to the search for biomarkers in autism spectrum disorder. *Curr. Opin. Neurol.* **29,** 123–9 (2016).
72. Birnbaum, R. & Weinberger, D. R. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat. Rev. Neurosci.* **18,** 727–740 (2017).
73. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16,** 441–58 (2015).
74. Kajiwara, Y. *et al.* Extensive proteomic screening identifies the obesity-related NYGGF4 protein as a novel LRP-interactor, showing reduced expression in early Alzheimer's disease. *Mol. Neurodegener.* **5,** 1 (2010).
75. Castellanos, F. X. & Tannock, R. Neuroscience of attention-deficit/hyperactivity disorder: the search for endophenotypes. *Nat. Rev. Neurosci.* **3,** 617–628 (2002).
76. Kuntsi, J. *et al.* Separation of Cognitive Impairments in Attention-Deficit/Hyperactivity Disorder Into 2 Familial Factors. *Arch. Gen. Psychiatry* **67,** 1159 (2010).

77. Kuntsi, J. *et al.* Genetic analysis of reaction time variability: room for improvement? *Psychol. Med.* **43,** 1323–1333 (2013).
78. Tamm, L. *et al.* Reaction Time Variability in ADHD: A Review. *Neurotherapeutics* **9,** 500–508 (2012).
79. Davies, G. *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol. Psychiatry* **21,** 758–767 (2016).
80. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
81. Perneger, T. V. What's wrong with Bonferroni adjustments. *BMJ* **316,** 1236–8 (1998).
82. Brett, M., Penny, W. & Kiebel, S. An Introduction to Random Field Theory. (2003).
83. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15,** 1–25 (2002).
84. Mishra, P., Törönen, P., Leino, Y. & Holm, L. Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics* **30,** 2747–2756 (2014).
85. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
86. Seshadri, S. *et al.* Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham study. *BMC Med. Genet.* **8,** S15 (2007).
87. Lenzenweger, M. F. Thinking clearly about the endophenotype–intermediate phenotype–biomarker distinctions in developmental psychopathology research. *Dev. Psychopathol.* **25,** 1347–1357 (2013).
88. Gottesman, I. I. & Gould, T. D. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *Am. J. Psychiatry* **160,** 636–645 (2003).
89. Bernal-Rusiel, J. L. *et al.* Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *Neuroimage* **66,** 249–60 (2013).
90. Brouwer, R. M. *et al.* Genetic influences on individual differences in longitudinal changes in global and subcortical brain volumes: Results of the ENIGMA plasticity working group. *Hum. Brain*

Mapp. **38,** 4444–4458 (2017).

91. Hoth, K. F. *et al.* Link between change in cognition and left ventricular function following cardiac resynchronization therapy. *J. Cardiopulm. Rehabil. Prev.* **30,** 401–8 (2010).
92. Carmichael, O. T. *et al.* Cerebral Ventricular Changes Associated With Transitions Between Normal Cognitive Function, Mild Cognitive Impairment, and Dementia. *Alzheimer Dis. Assoc. Disord.* **21,** 14–24 (2007).
93. Nestor, S. M. *et al.* Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain* **131,** 2443–54 (2008).
94. Thompson, P. M. *et al.* Mapping hippocampal and ventricular change in Alzheimer disease. (2004). doi:10.1016/j.neuroimage.2004.03.040
95. Mak, E. *et al.* Longitudinal whole-brain atrophy and ventricular enlargement in nondemented Parkinson's disease. *Neurobiol. Aging* **55,** 78–90 (2017).
96. Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Res.* **44,** 6046–54 (2016).
97. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93,** 779–97 (2013).
98. Lista, S., Garaci, F. G., Toschi, N. & Hampel, H. Imaging epigenetics in Alzheimer's disease. *Curr. Pharm. Des.* **19,** 6393–415 (2013).
99. Hampton, T. Imaging Epigenetics in the Human Brain. *JAMA* **316,** 1349 (2016).
100. Yan, J. *et al.* Identification of Discriminative imaging proteomics associations in Alzheimer's Disease via a novel sparse correlation model. *Biocomputing 2017* **22,** 94–104 (WORLD SCIENTIFIC, 2017).
101. Nazeri, A. *et al.* Imaging proteomics for diagnosis, monitoring and prediction of Alzheimer's disease. *Neuroimage* **102,** 657–665 (2014).
102. Sui, J., Yu, Q., He, H., Pearlson, G. D. & Calhoun, V. D. A selective review of multimodal fusion methods in schizophrenia. *Front. Hum. Neurosci.* **6,** 27 (2012).
103. Zhu, D. *et al.* Fusing DTI and fMRI data: a survey of methods and

applications. *Neuroimage* **102 Pt 1,** 184–91 (2014).
104. Schouten, T. M. *et al.* Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage. Clin.* **11,** 46–51 (2016).
105. Liu, K. *et al.* Transcriptome-Guided Imaging Genetic Analysis via a Novel Sparse CCA Algorithm. in 220–229 (Springer, Cham, 2017). doi:10.1007/978-3-319-67675-3_20
106. Meyer-Lindenberg, A. The future of fMRI and genetics research. *Neuroimage* **62,** 1286–1292 (2012).
107. Hyde, L. W., Bogdan, R. & Hariri, A. R. Understanding risk for psychopathology through imaging gene-environment interactions. *Trends Cogn. Sci.* **15,** 417–27 (2011).
108. Dunlop, B. W. & Mayberg, H. S. Neuroimaging-based biomarkers for treatment selection in major depressive disorder. *Dialogues Clin. Neurosci.* **16,** 479–90 (2014).
109. Huibers, M. J. H. *et al.* Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLoS One* **10,** e0140771 (2015).

ANNEX

Annex I. PhD Portfolio

About the author

Natàlia Vilor-Tejedor was born in Barcelona in April 1988. She received a BSc degree in Mathematics and Applied Statistics from the Universitat Autònoma de Barcelona (UAB), in June 2012. She also received a MSc degree in Omics Data Analysis from the University of Vic-Universitat Central de Catalunya in February 2014. In 2015, she received a fellowship to carry out her PhD at Centre for Research in Environmental Epidemiology (CREAL), now Barcelona Institute for Global health (ISGlobal). As part of her PhD training, she did a 4-month stay at the Erasmus Medical Center-University of Rotterdam (August 2017-December 2017). She also did a two-week stay at the University of Basque Country-Departament of Mathematics (October, 2016) and a three-week stay at the University of Harvard- Harvard T.H. Chan School of Public Health and Harvard Medical School (March-April, 2017). A complete summary of the research activity of the author during the thesis is provided below.

List of Publications

Vilor-Tejedor N, Ikram MA, Roshchupkin G, Niessen WJ, Alemany S, Adams HH. *Genome-wide genetic risk variants for ADHD predict longitudinal changes of ventricle structures in a population-based sample* (in preparation).

Vilor-Tejedor N, Ikram MA, Roshchupkin GV, Cáceres A, Alemany S, Vernooij MW, Niessen WJ, van Duijn CM, Bustamante M, Pujol J, Sunyer J, Adams HH, González JR.

*Independent Multifactorial Association Analysis to analyze multiblock data in Imaging Genetics (*in preparation*).*

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Pujol J, Sunyer J, González JR. *Strategies for integrative analysis in Imaging Genetic studies. (under review since 6th December, 2017. Neuroscience and biobehavioral reviews journal).*

Mortamais M Pujol J, Macià D, Martínez-Vilavella G, Fenoll R, Reynes C, Sabatier R, Rivas I, Forns J, **Vilor-Tejedor N**, Alemany S, Alvarez-Pedrerol M, Sunyer J. *Effects of prenatal exposure to particulate matter air pollution on lateral ventricles, corpus callosum and behavioral problems in children.(under review, JAMA Psychiatry).*

Cacheiro P, Lorenzo-Arribas A, Alvares D, Barrio I, Bofill-Roig M, Branco M, Lázaro E, Pérez Haro MJ, Gómez-Mateu M, **Vilor-Tejedor N.** *Reproducibility, visibility and diversity: current trends in Biostatistics viewed by young researchers.(under review, Bulletin of Statistics and Operations Research).*

López-Vicente M, Ribas Fitó N, **Vilor-Tejedor N,** Garcia-Esteban R, Fernández-Barrés S, Dadvand P, Murcia M, Rebagliato M, Ibarluzea J, Lertxundi A, Fernández-Somoano A, Tardón A, Romaguera D, Vrijheid M, Sunyer J, Júlvez J. *Prenatal omega-6/omega- 3 ratio and attention deficit and hyperactivity disorder symptoms in children: a population-based longitudinal study.(under review. The American Journal of Clinical Nutrition.)*

Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Mortamais M, Pujol J, Sunyer J, González JR. *Sparse multifactorial analysis reveals the role of cerebellar tissue volumes and molecular processes in ADHD dimensions. (Under review,*

since 5th October 2017. International Journal of Methods in Psychiatric Research).

Alemany S, **Vilor-Tejedor N**, García-Esteban R, Bustamante M, Dadvand P, Mortamais M, Forns J, van Drooge JV, Alvárez-Pedrerol M, Rivas I, Querol X, Pujol J, Sunyer J. *Traffic air pollution, APOE ε4 status and neurodevelopment in scholar children (under review Environmental Health Perspectives.)*

Vilor-Tejedor N, Cáceres A, Pujol J, Sunyer J, González JR. *Imaging genetics in attention-deficit/hyperactivity disorder and related neurodevelopmental domains: state of the art.* Brain Imaging Behav. 2016 Dec 15. [Epub ahead of print] Review. PubMed PMID: 27981420.

Alemany S, **Vilor-Tejedor N**, Bustamante M, Álvarez-Pedrerol M, Rivas I, Forns J, Querol X, Pujol J, Sunyer J. *Interaction between airborne copper exposure and ATP7B polymorphisms on inattentiveness in scholar children.* Int J Hyg Environ Health. 2016 Oct 22. pii: S1438-4639(16)30257-7. doi: 10.1016/j.ijheh.2016.10.010. [Epub ahead of print] PubMed PMID: 28029585.

Gomez-Mateu M, Lorenzo-Arribas A, Bofill-Roig M, **Vilor-Tejedor N,** Barrio I, Espasandin-Dominguez J, Guler I, Cacheiro P, Aguirre U, Pérez-Álvarez N. *Big Data in Biomedical Research. Perspectives from the Biostatnet-CRM Workshop.* BEIO 2016; 32(3): 257-277.

Horikoshi M, Beaumont RN, Day FR, ..., **Vilor-Tejedor N**,..., Timpson NJ, Perry JR, Evans DM, McCarthy MI, Freathy RM. *Genome-wide associations for birth weight and correlations with adult disease.* Nature. 2016 Oct 13;538(7624):248-252. doi:

10.1038/nature19806. PubMed PMID: 27680694; PubMed Central PMCID: PMC5164934.

Alemany S, **Vilor-Tejedor N**, Bustamante M, Pujol J, Macià D, Martínez-Vilavella G, Fenoll R, Alvárez-Pedrerol M, Forns J, Júlvez J, Suades-González E, Llop S, Rebagliato M, Sunyer J. A *Genome-Wide Association Study of Attention Function in a Population-Based Sample of Children*. PLoS One. 2016 Sep 22;11(9):e0163048. doi: 10.1371/journal.pone.0163048. PubMed PMID: 27656889; PubMed Central PMCID: PMC5033492.

Vilor-Tejedor N, Alemany S, Forns J, Cáceres A, Murcia M, Macià D, Pujol J, Sunyer J, González JR. *Assessment of Susceptibility Risk Factors for ADHD in Imaging Genetic Studies*. J Atten Disord. 2016 Aug 17. pii: 1087054716664408. [Epub ahead of print] PubMed PMID: 27535943.

Middeldorp CM, Hammerschlag AR, Ouwens KG, Groen-Blokhuis MM, St Pourcain B, Greven CU, Pappa I, Tiesler CM, Ang W, Nolte IM, **Vilor-Tejedor N**,..., Tiemeier H, Posthuma D, Boomsma DI. *A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts*. J Am Acad Child Adolesc Psychiatry. 2016 Oct;55(10):896-905.e6. doi: 10.1016/j.jaac.2016.05.025. PubMed PMID: 27663945; PubMed Central PMCID:PMC5068552.

Marinelli M, Pappa I, Bustamante M, Bonilla C, Suarez A, Tiesler CM, **Vilor-Tejedor N**,..., Smith GD, Estarlich M, Heinrich J, Räikkönen K, Vrijkotte TG,Tiemeier H, Sunyer J. *Heritability and Genome-Wide Association Analyses of Sleep Duration in Children: The EAGLE Consortium*. Sleep. 2016 Aug 19. pii: sp-00625-15. [Epub ahead of print] PubMed PMID: 27568811.

Bustamante M, Standl M, Bassat Q, **Vilor-Tejedor N**,..., Moll HA, Heinrich J, Estivill X, Sunyer J. *A genome-wide association meta-analysis of diarrhoeal disease in young children identifies FUT2 locus and provides plausible biological pathways*. Hum Mol Genet. 2016 Aug 23. pii: ddw264. [Epub ahead of print] PubMed PMID: 27559109.

Felix JF, Bradfield JP, Monnereau C,..., **Vilor-Tejedor N**,..., Timpson NJ, Grant SF, Jaddoe VW; Early Growth Genetics (EGG)Consortium; Bone Mineral Density in Childhood Study BMDCS. *Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index*. Hum Mol Genet. 2016 Jan 15;25(2):389-403. doi: 0.1093/hmg/ddv472. Epub 2015 Nov 24. PubMed PMID: 26604143.

Vilor-Tejedor N, Gonzalez JR, Calle ML. *Efficient and powerful method for combining p-values in Genome-wide Association Studies*. IEEE/ACM Trans Comput Biol Bioinform. 2015 Dec 22. doi: 10.1109/TCBB.2015.2509977. [Epub ahead of print]. PubMed PMID: 28055892.

Paternoster L, ..., **Vilor-Tejedor N**, ..., Sunyer J,..., Weidinger S. *Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis*. Nat Genet 2015; 47(12): 1449-56.

Alemany S, Ribasés M, **Vilor-Tejedor N**, Bustamante M, Sánchez-Mora C, Bosch R, Richarte V, Cormand B, Casas M, Ramos-Quiroga JA, Sunyer J. *New suggestive genetic loci and biological pathways for attention function in adult attention-deficit/hyperactivity disorder*. Am J Med Genet B Neuropsychiatr Genet. 2015 Jul 14. doi: 10.1002/ajmg.b.32341. [Epub ahead of print] PubMed PMID: 26174813.

Vilor-Tejedor N and Calle ML. *Global adaptive rank truncated product method for gene-set analysis in association studies*. Biom J. 2014 Sep;56(5):901-11. doi: 10.1002/bimj.201300192. PubMed PMID: 25082012.

Summary of PhD Training and Teaching

Oral Communications

"*Statistical Methods in Genetics*". 3rd Young Spanish Biometric Conference, 18-19th January 2018, Bilbao, Spain.

"*Independent multifactorial analysis to analyze multiblock data: application in Imaging Genetic studies*". Spanish Biometric Conference, 12-15th September 2017, Sevilla, Spain.

"*Multiple correspondence analysis to construct an overall score of attention-deficit/hyperactivity disorder symptoms*".
3rd BIOSTATNET General Meeting: facing biostatistical research challenges with international projection, 19-21st January, 2017, Santiago de Compostela, Spain.

"*Assessment of Susceptibility Risk Factors for neurodevelopmental domains in Imaging Genetic Studies*". ISGlobal PhD Symposium, 28th November 2016. Barcelona, Spain.

"*Assessment of imaging genetic correlates for Attention-deficit/Hyperactivity Disorder*". 2nd Young Spanish Biometric Conference. 8-9th September 2016, Barcelona, Spain.

"*Novel analytical insights to evaluate Attention-deficit/Hyperactivity Disorder symptoms in Imaging Genetic*

studies". Europe's Young Researchers Conference on Environmental Epidemiology. 1-3rd November 2015, Utrecht, The Netherlands

"*Independent Multifactorial association Analysis to analyze multiblock data: application in an Imaging Genetic study of AttentionDeficit/Hyperactivity Disorder*", Conferencia Española de Biometría. 21-25th September 2015, Bilbao, Spain

"*Statistical methods in Imaging Genetics*". 1st Young Spanish Biometric Conference. 19-20th January 2015, Valencia, Spain.

"*Integrating Genome-wide association results and structural magnetic resonance data for childhood Attention-deficit hyperactivity disorder symptoms*". Systems Neuroscience Symposium. 17th December 2014, Barcelona, Spain.

"*Efficient and powerful testing for gene set analysis in Genome-Wide Association Studies*". Congreso Jovenes Investigadores en Bioestadística y Diseño de Experimentos. 21-22nd July 2014, Pamplona, Spain

Poster Communications

Vilor-Tejedor N, Alemany S, Caceres A, Bustamante M, Sunyer J, González JR. "Independent Multifactorial Analysis to analyze multiblock data: application to Imaging Genetic studies". IV Bioinformatics and Genomics Symposium. 20th December 2017, Barcelona, Spain.

Vilor-Tejedor N, Caceres A, Alemany S, Bustamante M, Sunyer J, González JR. "Independent Multifactorial Analysis to analyze

multiblock data: application to Imaging Genetic studies". IV ISGlobal PhD symposium. 28th November, 2017, Barcelona, Spain.

Vilor-Tejedor N, Caceres A, Alemany S, Bustamante M, Sunyer J, González JR. "Independent Multifactorial Analysis to analyze multiblock data: application to Imaging Genetic studies". V Bioinformatics Student Symposium, 12nd May 2017, Barcelona, Spain.

Vilor-Tejedor N, Sunyer J, González JR. "Imaging Genetics in childhood Attention-Deficit/Hyperactivity Disorder". Annual scientific meeting - Proyecto INMA 2016. 27-28th October 2016, Barcelona, Spain.

Vilor-Tejedor N, Sunyer J, González JR. "Novel analytical strategies to assess potential associations in Imaging Genetic studies of childhood Attention-Deficit/Hyperactivity Disorder symptoms and neurocognition". 2nd ISGlobal PhD Symposium. 4th November 2015, Barcelona, Spain.

Alemany S, **Vilor-Tejedor N**, Bustamante M, Pujol J, Macià D, Álvarez-Pedrerol M, Sunyer J. "Gene-environment interaction between copper-transportin ATPase genes and copper exposure in childhood inattentiveness". Europe's Young Researchers Conference on Environmental Epidemiology. 1-3rd November, 2015, Utrecht, The Netherlands.

Vilor-Tejedor N, Alemany S, Macia D, Bustamante M, Pujol J, Sunyer J, González JR. "Integrating Genome-wide association results and structural magnetic resonance data for childhood Attention-deficit hyperactivity disorder symptoms." Third DCEXS Symposium. New insights in Genetics and Neuroscience research, 26th November 2014, Barcelona, Spain.

Vilor-Tejedor N, Sunyer J, González JR. "Integration of Genomics and Neuroimaging data in Cognitive Development". 1st ISGlobal PhD Symposium. 13rd November 2014, Barcelona, Spain.

Vilor-Tejedor N, González JR, Calle M. "*Gene Set Analysis for genetic association studies using the extreme value theory*". International Biometric Conference. 6-11th July 2014. Florence, Italy.

Training Seminars and Courses (Offered by the Author)

Programming for Bioinformatics. Subject from Master's degree in Omics Data Analysis – University of Vic – Central University of Catalonia (UVIC/UCC). Course 2017-2018.

Introduction to R – 3rd edition. Seminar for researchers, statisticians and students of the ISGlobal (duration: 2 h), 31st May 2017, Pompeu Fabra University, Barcelona, Spain.

Computational Tools for systems Medicine organized by the OpenMultiMed CA15120 (Open Multiscale Systems Medicine, COST Action CA15120) (duration: 5 hrs), 21-23rd, 2017, Porto, Portugal.

Programming for Bioinformatics. Subject from Master's degree in Omics Data Analysis – University of Vic (UVIC/UCC). Course 2016-2017.

Introduction to R – 2nd edition. Seminar for researchers, statisticians and students of the ISGlobal (duration: 2 h), 11st March 2015, Centre for research in environmental Epidemiology, Barcelona, Spain.

Introduction to R – 1st edition. Seminar for researchers, statisticians and students of the ISGlobal (duration: 2 h), 22nd May 2014, Center for Research in Environmental Epidemiology, Barcelona, Spain.

Training Seminars and Courses (Taken by the Author)

14th Course on SNP's and Human Diseases. November 2017. Molecular Medicine Postgraduate School. Erasmus Medical Center, Rotterdam.

10 keys to creating great visual aids for scientific presentations. November 2016. PRBB Intervals. Barcelona Biomedical Research Park, Barcelona.

Introduction to Imaging Genetics. June 2016. Human Brain Mapping Conference, Geneve.

Effective Writing for Biomedical Professionals. November 2015. Oxford University, Oxford, UK.

Statistical Learning. January 2015. Standford University (online).

Research Integrity in Biomedical Sciences. December 2014. Epigeum Online Course System. Oxford University Press.

MetaAnalysis using R. June 2014. Universitat Politecnica Catalunya, Barcelona.

Statistical Analysis of fMRI Data. April 2014. John Hopking University, (online).

Make your Research viral. March 2014. PRBB Intervals. Barcelona Biomedical Research Park, Barcelona.

Statistical Reasoning for Public Health. March 2014 (online).

Metodología de las publicaciones biomédicas. January 2014. PRBB Intervals. Barcelona Biomedical Research Park, Barcelona.

Topics on the analysis of cohorts. November 2013. Dipartimento di Epidemiologia. Roma, Italy.

Awards and Fellowships

2017. Grant for a short Term Scientific Mission (STSMs) - Open Multiscale Systems Medicine OpenMultiMed COST Action 15120. Call 2017.

2017. Grant for abroad stay to complete a Ph.D with international mention – Centro de Investigación Biomédica en Red (CIBER) – Call 2017.

2017. Student Travel Award Winner – Spanish Biometric Society - XVI Spanish Biometric Conference for the abstract presentation: "Independent Multifactorial Association Analysis to analyze multiblock data: application in an Imaging Genetic study of Attention-Deficit/Hyperactivity Disorder ", XV Spanish Biometric Conference, 12-15th September, 2017, Sevilla, Spain.

2017. Student Travel Grant and registration fee waivers for a short formative stay at Harvard T.H. Chan School of Public Health. Agency for Management of University and Research Grants

(AGAUR) co-financed with the European Social Fund. 22nd-07th April, 2017. Harvard University, Boston, MA.

2017. Student Travel Grant and registration fee waivers. Agency for Management of University and Research Grants (AGAUR) co-financed with the European Social Fund. 3rd General Meeting of the Spanish National Network of Biostatistics, 19-21st, January, 2017. Santiago de Compostela, Spain.

2016. Scholarship for a short formative stay in the framework of the BIOSTATNET Youth Researchers Program call 2016 - Spanish National Network of Biostatistics, 10-21st, October, 2016. University of the Basque Country (UPV/EHU), Bilbao, Spain.

2016. Student Travel Grant and registration fee waivers – Center for Research in Environmental Epidemiology - ISEE Europe's Young Researchers Conference on Environmental Epidemiology for the abstract presentation: Novel analytical insights to evaluate Attention-deficit/Hyperactivity Disorder symptoms in Imaging Genetic studies, 1-3rd November, 2015, Utrecht, Netherlands.

2015. Student Travel Grant and registration fee waivers - Agency for Management of University and Research Grants (AGAUR) co-financed with the European Social Fund. - Effective Writing for Biomedical Professionals course, 18-21st, November 2015, Oxford, UK.

2015. Student Travel Grant and registration fee waivers - Agency for Management of University and Research Grants (AGAUR) co-financed with the European Social Fund. - 18th International Conference on Medical Image Computing and Computed Assisted Interventions, 10th October, 2015, Munich, Germany.

2015. Student Travel Award Winner – Spanish Biometric Society - XV Spanish Biometric Conference for the abstract presentation: "Independent Multifactorial Association Analysis to analyze multiblock data: application in an Imaging Genetic study of Attention-Deficit/Hyperactivity Disorder ", XV Spanish Biometric Conference, 21-25th September, 2015, Bilbao, Spain.

2015. FI-DGR predoctoral fellowship – Agency for Management of University and Research Grants (AGAUR) co-financed with the European Social Fund. Register number: FI_B 00636 (Duration 3 years).

2014. Speedy Oral Presentation Award - Centre for Genomic Regulation (CRG), Cellular and Systems Neurobiology lab & Neurosciences Research Program-Hospital del Mar Medical Research Institute (IMIM). "Integrating Genome-wide association results and structural magnetic resonance data for childhood Attention-deficit/hyperactivity disorder symptoms", 17th December, 2014, Barcelona, Spain.

2014. Lecture Award in the National Young Congress on Biostatistics and Experimental Designs - Spanish National Network of Biostatistics "Efficient and Powerful Testing for Gene Set Analysis in Genome-Wide Association Studies", 22nd July, 2014, Pamplona, Spain.

Annex II. Protocols

BREATHE project. Genotyping protocol

DNA extraction

Two thousand six hundred and forty three saliva samples from children were collected using the Oragene DNA OG-500kit (DNAGenotek). DNA extraction was done following manufacturer's instructions with minor modifications. DNA was eluted in 500-800 ul of TE and quantified using a NanoDrop 1000 UV-Vis Spectrophotometer (ThermoScientific). Integrity of DNA was checked in a subset of samples by running a 1% agarose gel. DNAs were aliquoted in two micronic plates: one was stored at -20C and the other one at -80C at Center for research in environmental epidemiology's biorepository. After elimination of duplicates and samples not properly coded we ended up with 2,492 unique DNAs (approximately 85.5% of BREATHE children had DNA).

Selection of samples

From the initial biorepository (N=2,492) we filtered low quality DNAs, children that did not have neuropsychological test, children with a non Caucasian descent origin and not born in Spain and children whose parents were not born in Europe. We also filtered adopted children and tried to keep only one sibling per family (based on address, school, parental education, parental jobs and number of siblings at birth information). Finally, in the analysis we ended up with 1,778 BREATHE samples and 20 BREATHE duplicates. This represents around 72% of the biorepository samples and 62% of the children in BREATHE.

Details of Lab processing

A total of 1,798 samples from BREATHE project (1,778 samples and 20 duplicates) and 74 CEU HapMap controls (54 CEU trios and 20 duplicates) were genotyped using the HumanCore BeadChip WG-330-1101 array from Illumina at the Spanish National Genotyping Center (CEGEN) at the Spanish National Cancer Research Centre (CNIO). Before genotyping DNAs were quantified using Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies). Genotype calling was done with the GeneTrain 2.0 algorithm (Illumina) with a default threshold of 0.15 and based on HapMap clusters preimplemented in GenomeStudio software. HapMap callings were consistent with genotypes in the public databases and Mendelian errors were below 0.001%. Genetic markers are reported on the Human Reference Genome hg19 (b37) and on the + strand.

BREATHE project. Quality control protocol

Definitions of Quality Measures

Details of quality control (QC) performing from BREATHE genome-wide data can be found in the following sections. This procedure is an adaptation from *Anderson et al. (2011).* We will begin by performing sample quality control, including identification of individuals with outlying missing genotype or heterozygosity rates, identification of individuals with discordant sex information, identification of duplicated or related individuals and identification of individuals of divergent ancestry. We will then perform genotype quality control including calculation of call rates, analysis of minor allele frequency (MAF) and deviation from Hardy-Weinberg equilibrium (HWE). The quality control analysis is performed with the PLINK software (*Purcell et al. 2007*).

Performance of Quality Control

Before starting the quality control procedure we will determine our working directory path that was:

```
/DATA/BREATHE_GWAS/GWAS/QC_NV/
```

Also, we will need to create the genome-wide association binary ped files. Hence, we make a binary ped file (*.bed). This will store the pedigree/phenotype information in separate file (*.fam) and create an extended MAP file (*.bim) which contains information about the allele names, which would otherwise be lost in the *.bed file. To create these files we use the command:

```
plink --file
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS --out
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS --make-bed -noweb
```

We keep binary files into:
`/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/`

Performance of Sample quality control

i. CALL RATE:

At the shell prompt, type:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS_upd_IDs --missing --out full_data_table_GS -
noweb
```

This command creates "`full_data_table_GS.imiss`" and "`full_data_table_GS.lmiss`" files, where the fourth column in the file "full_data_table_GS.imiss" (N_MISS) denotes the number of missing SNPs and the sixth column (F_MISS) denotes the

proportion of missing SNPs per individual. The missing call rate per sample is an informative indicator that identifies the samples with missing call rates over a predefined percentage of significant level. We decided to exclude individuals with a genotype failure rate >0.03 (call rate<97%).

ii. HETEROZYGOSITY:

The distribution of mean heterozygosity across individuals is inspected to detect an excessive or reduced proportion of heterozygote genotypes that suggest a contamination or inbreeding of DNA sample, respectively. At the shell prompt type:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS_upd_IDs --het --out full_data_table_GS -
noweb
```

This command will create the file "`full_data_table_GS.het`", in which the third column denotes the observed number of homozygous genotypes [O(Hom)] and the fifth column denotes the number of non-missing genotypes [N(NM)] per individual. Then, we can calculate the observed heterozygosity rate per individual using the formula: Het = (N(NM) − O(Hom))/N(NM).
The estimate of Heterozygosity (F) can sometimes be negative. Often this will just reflect random sampling error, but a result that is strongly negative (i.e. an individual has *fewer* homozygotes than one would expect by chance at the genome-wide level) can reflect other factors, e.g. sample contamination events perhaps. We filter samples with +/- 4 SD of the heterozigosity mean rate

iii. IDENTIFICATION OF INDIVIDUALS WITH DISCORDANT SEX INFORMATION

To detect discrepancies between genotype information and sex information we calculate homozygosity across the X-chromosome. This procedure is expected to detect differences between sex information because males have only one copy of the X chromosome and it implies that they cannot be heterozygous for any marker, contrary to females (except for the pseudoautosomic regions). Hence, one expects male samples to have a homozygosity rate around of 1 while female samples will have, approximately, a homozygosity rate less of 0.2. We chose to excluded samples which do not satisfy this criteria. At the shell prompt, type:

```
plink --file
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS_upd_IDs --check-sex --out full_data_table_GS
--noweb
```

This option uses X chromosome data to determine sex (i.e. based on heterozygosity rates) and flags individuals for whom the reported sex in the PED file does not match the estimated sex (given genomic data). We excluded of individuals with discordant sex information

iv. IDENTIFICATION OF DUPLICATED OR RELATED INDIVIDUALS.

It is also important to estimate the unexpected relatedness between samples. We use the IBD (identical by descendent) measure. A DNA segment is IBD in two or more individuals if they have inherited it from a common ancestor without recombination, that is, the segment has the same ancestral origin in these individuals. Note In a homogeneous sample**,** it is possible to calculate genome-wide IBD given IBS (identical by state)

information, as long as a large number of SNPs are available (probably 1,000 independent SNPs at a bare minimum; ideally 100K or more)

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS --genome --out full_data_table_GS --noweb
```

The procedure estimates the three identity-by-descent (IBD) probabilities of sharing 0, 1 or 2 alleles that are identical by descent. In addition, the PI_HAT measure is calculated.
We expect to get:

- A PI_HAT=1 for identical twins and duplicates, because they are 100% IBD.
- A PI_HAT=0.5 for first-degree relatives that are defined as 50% IBD.
- A PI_HAT=0.25 for second-degree relatives that are defined as 25% IBD.
- A PI_HAT = 0.125 for third-degree relatives that are defined as 12.5% IBD.

Based on these criteria, we excluded individuals with IBD > 0.185

v. IDENTIFICATION OF INDIVIDUALS OF DIVERGENT ANCESTRY.

To identify individuals of divergent ancestry we use Principal Component Analysis (PCA) described in more detail by *Patterson et al. (2006)* on IBD (identical by descendent) measure estimated in the previous section. The two first components show a homogeneous cluster indicating that it isn't necessary to exclude any sample based on Principal Component Analysis. In addition, we can also observe and homogeneous behaviour between female and male samples, and more importantly BREATHE subjects do not differ much from CEU HapMap samples, however HapMap

subjects tend to be located in the extreme of the cluster (they are North European descendent and not Spanish). Hence, we don't exclude any sample based on Principal Component Analysis.

At the shell prompt, type:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/full_data_
table_GS --cluster --read-genome
full_data_table_GS.genome --mds-plot 4 --K 2 --out
strat -noweb
```

Performance of genotype data quality control.

i. IDENTIFICATION OF ALL MARKERS WITH AN EXCESSIVE MISSING DATE RATE (CALL RATE)

i.1) Explore genotyping/missingness in the data.
To calculate the missing genotype rate for each marker, type:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/clean-
GWAS-data --missing --out clean-GWAS-data --noweb
```

Results of this analysis can be found in `'clean-inds-GWA-data.lmiss'`.
We can also plot a histogram of the missing genotype rate to identify a threshold for extreme genotype failure rate. We chose a call-rate threshold of 5%.

i.2) Pruning the data based on advised >95%.
Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/clean-
GWAS-data --geno 0.05 --make-bed --out
```

```
/DATA/BREATHE_GWAS/GWAS/QC_NV/QC_data/BREATHE_QC1_CR
-noweb
```

ii. MINOR ALLELE FREQUENCY (MAF)

ii.1) Explore MAF in the data.

To explore MAF for each marker, type:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/clean-
GWAS-data --freq --out clean-GWAS-data -noweb
```

It creates a file called `clean-GWAS-data.frq` with 298,930 SNPs.

ii.2) Pruning the data based on advised >1%.

Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency).

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/QC_data/BREATHE_QC1_CR
--maf 0.01 --make-bed --out
/DATA/BREATHE_GWAS/GWAS/QC_NV/QC_data/BREATHE_QC1_CR_
MAF - noweb
```

iii. HARDY WEINBERG EQUILIBRIUM (HWE)

In order to check the allele frequency in a population, we can use the Hardy- Weinberg equation.

iii.1) Explore HWE in the data.

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/Binary_files/clean-
GWAS-data --hardy --out clean-GWAS-data --noweb
```

iii.2) Pruning the data based on advised at least p>10-6.

To exclude markers that failure the Hardy-Weinberg test at a specified significance threshold, use the option:

```
plink --bfile
/DATA/BREATHE_GWAS/GWAS/QC_NV/QC_data/BREATHE_QC1_CR_
MAF --hwe 0.000001 --make-bed --out
/DATA/BREATHE_GWAS/GWAS/QC_NV/QC_data/BREATHE_QC1_CR_
MAF_HWE - noweb
```

iv. DESCRIPTION OF THE FINAL DATASET OF GENETIC MARKERS.

The final genetic data set consists of 1,667 subjects and 246,103 SNPs in b37 and + strand. Mean MAF is 0.275.

BREATHE project. Imputation analysis protocol

i.SETTING DATA FOR IMPUTATION

i.1) Create .gen and .sample files

```
gtool -P --ped BREATHE_QC_FINAL.ped --map
BREATHE_QC_FINAL.map --og BREATHE_QC_FINAL.gen --os
BREATHE_QC_FINAL.sample
```

i.2. Split the data into different chromosomes

```
for i in {1..26}
do
echo \BREATHE_QC_FINAL\"$i\"
o=\BREATHE_QC_FINAL\"$i\"
plink --bfile BREATHE_QC_FINAL --chr $i --recode --
out $o --noweb
gtool -P --ped $o.ped --map $o.map --og $o.gen --os
$o.sample
done
```

We refereed to chromosome 26 for Mitocondrial, to chromosome 25 for PAR regions and to chromosome 23 and 24 for chromosome X and Y.

i.3) Download 1000 genome release March 2012

Reference Panel: 1000 Genomes Phase I integrated variant set (b37, Mar.2012, Includes chrX updated 24 Aug 2012).

Link:

https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html

ii. IMPUTATION ANALYSIS (IMPUTE v2)

We developed several script to perform the imputation procedure.

ii.1) Autosomal Chromosomes

- First we perform prephasing and imputation of the whole genome. The procedure took 10 processors in screen mode as a default. Data is split into chromosome branches, so there is a total of 44 chromosomes.

  i.e file: `IMPUTED_chr43_5.phased.impute2` *contains the imputations for chromosome 22, first branch chunk 5.*

  The correct interpretation of the chunks are given in a file, where the real chromosome and coordinates are written in the last three columns and the numbering of the files is given by the first to columns.

- Second, we identified those chunks with less than 200 SNPs in the genotyped data to be merged with contiguous chunks, so resulting chunks have enough SNPs for imputation.

- Third, we evaluate if the different chunks run correctly.

- Fourth, we recomputed the prephasing and imputation of the previous merged chunks. The procedure took 6 processors in screen mode as a default.
- Finally, we merged the final results into chromosomes.

ii.2) Sexual Chromosomes and Pseudoautosomal regions

Prephasing and imputation were undertaken separately for males and females. For females, the same protocol as for autosomes is used. For males, the pseudo-autosomal part of the X chromosome is removed.

iii. QUALITY CONTROL IMPUTATION ANALYSIS

We additionally developed a script to perform the quality control for the genotyped and imputed markers. The implemented procedure also builds the summary tables.

BREATHE project. Extraction of genotyped and imputed variants

Genotyped variants

i. PREPARE A LIST OF SNPS IN A *.txt* FORMAT

Here an example for the mysnps.txt:

rs2056974
rs2473316
rs12740705
rs3170633

Note:
No TAB or space between snps are required.

ii. ENTER THE PATH FOR RESEARCHERS:

```
########in the console

cd /DATA/BREATHE_GWAS/DATA_FOR_RESEARCHER/
```

iii. CREATE A FOLDER FOR SAVING THE STUDY DATA.

An example:

```
########in the console

 mkdir  NameResearcher_SurnameResearcher
```

iv. COPY THE LIST OF SNPS IN THE RESEARCHERS DIRECTORY USING FILEZILLA OR WINSCP

v. ENTER IN THE STUDY RESEARCHER PATH

```
########in the console

cd
/DATA/BREATHE_GWAS/DATA_FOR_RESEARCHER/NameResearcher
_SurnameResearcher
```

vi. EXTRACT THE SNPS USING THE *mysnps.txt* FILE

```
########in the console

plink --bfile
/DATA/BREATHE_GWAS/GWAS/BD_FINAL_QC/BREATHE_QC_FINAL
--extract mysnps.txt --make-bed --out
snps_resarcher_DAYMONTHYEAR --noweb
```

Three files will be created with the make bed instruction:

```
snps_ resarcher_DAYMONTHYEAR.bed
snps_ resarcher_DAYMONTHYEAR.fam
snps_ resarcher_DAYMONTHYEAR.bim
```

.fam file includes:

Family ID
Individual ID
Paternal ID
Maternal ID
Sex (1=male; 2=female; other=unknown)
Phenotype

.bim file includes:

chromosome (1-22, 23=X, 24=Y, 25=XY, 26=Mit or 0 if unplaced)
rs# or snp identifier
Genetic distance (morgans)
Base-pair position (bp units): based on b36 (hg18)
A1 (minor allele)
A2 (major allele)

.bed file includes (binary file, genotype information)

A A G G A C
A A A G 0 0

vii. CREATE A FILE WITH SNP GENOTYPES RECODED IN TERMS OF ADDITIVE COMPONENTS

```
########in the console

plink --bfile snps_resarcher_DAYMONTHYEAR --recodeA --noweb --out snps_resarcher_ DAYMONTHYEAR _recoded –noweb
```

One file will be created:
"`snps_resarcher_DAYMONTHYEAR_recoded.raw`" with the

following variables:
"FID": family id
"IID": subject id (cohort_child_idnum)
"PAT": father id (empty)
"MAT": mother id (empty)
"SEX": sex is coded as 1=male; 2=female; other=unknown
"PHENOTYPE": -9 (empty)
"rsXXXXXXXX_T": 0, 1 or 2. Indicates number of T alleles (minor allele) the subject has.

Note: For alleles that have exactly 0.50 minor allele frequency, as for the second SNP in the example above, then which allele is labelled as minor will depend on which was first encountered in the PED file.

BREATHE project. Subcortical segmentation

i. PREPARATION

In your Linux environment, create a folder called "SUBJECTS" that contains the folders "input" and "output":

```
mkdir <path>/SUBJECTS
mkdir <path>/SUBJECTS/input SUBJECTS/output
```

Your input-folder should contain the scans of all your participants in a nii.gz-format. Make sure that each nii.gz-file is called `subj[number].nii.gz`

ii. SETUP FREESURFER

We followed all information found on this webpage:
https://surfer.nmr.mgh.harvard.edu/fswiki/QuickInstall

ii.1 Set up the variable FREESURFER_HOME

```
setenv FREESURFER_HOME <freesurfer_installation_direc
tory>/freesurfer
```

ii.2 Source the set up the Freesurfer script

```
source $FREESURFER_HOME/SetUpFreeSurfer.csh
```

ii.3 Set up the directory of subjects to work on:

```
setenv SUBJECTS_DIR <path>/enigma/input
```

iii. RUN PRE-PROCESSING

The next step is to run "recon–all" on the subjects in your input-folder (http://surfer.nmr.mgh.harvard.edu/fswiki/recon-all). In your input-folder, make a text-file containing a list of all your subjects.

Then, you can create the following script to run "recon-all" on multiple subjects:

```
#!/bin/bash
exec <List_subjects.txt
while read x; do
recon-all -i $x.nii.gz -s $x -all
mv $x/ <path>/SUBJECTS/output
done
```

To run this script, make it executable with the following command:

```
chmod u+x loop_recon_all
```

Then, run the script:

```
./loop_recon_all
```

Depending on the number of your scans and the processing speed of your computer, this script will take several days to finish (24 to 36 hours/subject).

IV. EXTRACT SUBCORTICAL VOLUMES

To extract subcortical volumes, run the following commands:

```
cd <path>/SUBJECTS/output
bash
echo
"SubjID,LLatVent,RLatVent,Lthal,Rthal,Lcaud,Rcaud,Lpu
t,Rput,Lpal,Rpal,Lhippo,Rhippo,Lamyg,Ramyg,Laccumb,Ra
ccumb,ICV" > SUBCvolumes.csv
for subj_id in `ls -d subj*`; do
printf "%s,"  "${subj_id}" >> SUBCvolumes.csv
for x in Left-Lateral-Ventricle Right-Lateral-
Ventricle Left-Thalamus-Proper Right-Thalamus-Proper
Left-Caudate Right-Caudate Left-Putamen Right-Putamen
Left-Pallidum Right-Pallidum Left-Hippocampus Right-
Hippocampus Left-Amygdala Right-Amygdala Left-
Accumbens-area Right-Accumbens-area; do
printf "%g," `grep  ${x} ${subj_id}/stats/aseg.stats
| awk '{print $4}'` >> SUBCvolumes.csv
done
printf "%g" `cat ${subj_id}/stats/aseg.stats | grep
IntraCranialVol | awk -F, '{print $4}'` >>
SUBCvolumes.csv
echo "" >> SUBCvolumes.csv
done
```

This will create a new file, called "SUBCvolumes.csv". It should contain a table with volumes (in mm^3).

Annex III. R Packages and Repositories

Package 'globalGSA'

November 17, 2014

Type Package

Title Global Gene-Set Analysis for Association Studies.

Version 2.0

Date 2014-11-17

Author Natalia Vilor-Tejedor, M.L. Calle

Maintainer Natalia Vilor-Tejedor <nvilor@creal.cat>

Description Implementation of four different Gene set analysis (GSA) algorithms for combining the individual pvalues of a set of genetic variats (SNPs) in a gene level pvalue. The implementation includes the selection of the best inheritance model for each SNP.

License GPL (>= 2)

R topics documented:

`globalGSA-package` *Gene-set analysis for combining p-values in a joint test of association between a phenotype and a set of genetic variants (SNPs). Previously, a global test for the best inheritance model of each SNP is performed.*

Description

This package implements four different Gene-set analysis (GSA) methods for combining individual p-values of a set of SNPs. Each method provides a p-value for a joint test of association between the phenotype and the specified set of genetic variants. The four implemented methods are: [1] the globalEVT method, [2] the globalARTP method, [4] the Fisher's method [5] the Simes' method. Since the SNPs in a set may follow different modes of inheritance, previously to the GSA, a global test for the best inheritance model (dominant, recessive, log-additive and co-dominant) is performed on every SNP. The permutational p-value of the best model is obtained.

Details

Package:	globalGSA
Type:	Package
Version:	2
Date:	2013-09-22
License:	GPL (>= 2)

Author(s)

Natalia Vilor, M.Luz Calle

Maintainer: nvilor@creal.cat

References

[1] Vilor-Tejedor N, Calle ML, Gonzalez JR. Efficient and powerful testing for gene set analysis applied to Genome-Wide association studies. (under submission)

[2] Vilor-Tejedor N and Calle ML. Global adaptive rank truncated product method for gene-set analysis in association studies. Biom. J. 2014; 56:901-911. doi: 10.1002/bimj.201300192

[3] Yu, K. Li, Q. Bergen, A.W. Pfeiffer, R.M. Rosenberg, P.S. Caporaso, N. Kraft, P. and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. Genet, Epidemiol. December; 33(8): 700-709.

[4] Fisher, R.A. (1925). Statistical Methods for Research Workers. ISBN 0-05-002170-2.

[5] Simes, R.J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 73, 751-754.

`globalARTP` *Global Adaptive Rank Truncated Product method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Adaptive Rank Truncated Product method [1] after a global test for the best mode of inheritance of every SNP [2]. The final gene-p-value is obtained from the permutational null distribution of the test statistic.

Usage

```
globalARTP(data, B, K, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

data	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
B	Number of permutations considered in the permutational procedure.
K	Integer that indicates the maximum truncation point.
gene_list	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
Gene	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
Trunkpoint	Considered truncation point.
Kopt	Optimal truncation point.
genevalue	gene-pvalue.

References

[1] Vilor-Tejedor N and Calle ML. Global adaptive rank truncated product method for gene-set analysis in association studies. Biom. J. 2014; 56:901-911. doi: 10.1002/bimj.201300192

[2] Yu, K. Li, Q. Bergen, A.W. Pfeiffer, R.M. Rosenberg, P.S. Caporaso, N. Kraft, P. and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. Genet, Epidemiol. December; 33(8): 700-709.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
```

```
# a direct association with the outcome:
data(data)
#globalARTP(data, B=1000, K=10, Gene="all", addit = FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans11)

# You can test:
globalARTP(data, B=1, K=10, Gene="all", addit = FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalARTP(data, B=1000, K=10, gene_list=gene_list, Gene="Gene1", addit = FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans1)

# You can test:
globalARTP(data, B=1, K=10, gene_list=gene_list, Gene="Gene1", addit = FALSE)
```

`globalEVT` *Global Adaptive Extreme Value Distribution method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using an Adaptive Extreme Value Distribution after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from

Usage

```
globalEVT(data, K, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial, LDinfo = NULL)
```

Arguments

`data`	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
`K`	Integer that indicates the maximum truncation point.
`gene_list`	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.

Gene	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.
LDinfo	Data frame containing the linkage disequilibrium between SNPs. By default, LDinfo=NULL.

Value

List with the following components:

Gene	Considered Gene.
Trunkpoint	Considered truncation point.
genevalue	gene-pvalue.

References

[1] Vilor-Tejedor N, Calle ML, Gonzalez JR. Efficient and powerful testing for gene set analysis applied to Genome-Wide association studies. (under submission)

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
globalEVT(data, K=10)
```

`globalFisher` *Global Fisher combination method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Fisher method [1] after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from the permutational null distribution of the test statistic

Usage

```
globalFisher(data, B, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

data	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
B	Number of permutations considered in the permutational procedure.
gene_list	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
Gene	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
genevalue	gene-pvalue.

References

[1] Fisher, R.A. (1925). Statistical Methods for Research Workers. ISBN 0-05-002170-2.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
#globalFisher(data, B=1000,  Gene="all", addit=FALSE)
```

```
# it may take some time,
# hence the result of this example is included:
data(ans21)

# You can test:
globalFisher(data, B=1,  Gene="all", addit=FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalFisher(data, B=1000, gene_list=gene_list, Gene="Gene1", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans2)

# You can test:
globalFisher(data, B=1, gene_list=gene_list, Gene="Gene1", addit=FALSE)
```

`globalSimes` *Global Simes' combination method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Simes method [1] after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from the permutational null distribution of the test statistic

Usage

```
globalSimes(data, B, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

`data`	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
`B`	Number of permutations considered in the permutational procedure.
`gene_list`	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
`Gene`	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we

	need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
genevalue	gene-pvalue.

References

[1] Simes, R.J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 73, 751-754.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
#globalSimes(data, B=1000,  Gene="all", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans31)

# You can test:
globalSimes(data, B=1,  Gene="all", addit=FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalSimes(data, B=1000, gene_list=gene_list, Gene="Gene1", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans3)

# You can test:
globalSimes(data, B=1, gene_list=gene_list, Gene="Gene1", addit=FALSE)
```

GitHub Profile

Search GitHub Pull requests Issues Marketplace Explore

Natalia Vilor-Tejedor

natvt8

Ph.D student: Institute for Global Health (ISGlobal) Assistant Lecturer: UVIC-UCC. Interests: BioStats, Neuroepidemiology, Imaging Genetics.

Barcelona

https://www.linkedin.com/in/na...

Overview Repositories 2 Stars 0 Followers 1 Following 2

Popular repositories

Customize your pinned repositories

globalGSA

Implementation of different Gene-set Analysis (GSA) algorithm for combining the individual pvalues of a set of genetic variants (SNPs) in a gene level p-value.

R

ICA-MFA

We present a novel multifactorial algorithm, referred as Independent Multifactor Association Analysis (IMFA-ICR), which uses Independent Component decomposition to derive relevant features from mul...

R

23 contributions in the last year

Contribution settings